Capitolo 1.

Il genoma umano e le tecnologie del DNA

Organizzazione molecolare e sequenze del genoma umano

Il genoma aploide umano è costituito da 25 specie molecolari DNAds:
→ 24 specie molecolari costituiscono il DNAds cromosomico nucleare (22 autosomi e 2 cromosomi sessuali X e Y). Nei cromosomi nucleari le molecole sono associate a proteine, le più abbondanti sono gli istoni che complessivamente hanno un peso uguale a quello del DNA nucleare. Il genoma nucleare aploide è costituito da circa $3{,}3 \times 10^9$ coppie di basi.
→ 1 specie molecolare di DNAds mitocondriale: un cromosoma circolare di 16.569b include 27 geni, 5-10 copie del cromosoma per mitocondrio, 1000-10.000 copie del cromosoma per cellula contenente mitocondri.

Sequenze del DNA nucleare:

DNA codificante proteine

DNA codificante ripetuto in tandem

rRNA, 5sRNA, tRNA e geni degli istoni varianti di replicazione.

DNA non codificante, costituito da sequenze ripetute e disperse nel genoma

Sequenze con un numero variabile di ripetizioni in tandem, localizzate su più loci
VNTR (Variable Number of Tandem Repeats) sequences.

Sequenze singole localizzate su più loci:
SINES (Short Interdispersed Nuclear Elements).
LINES (Long Interdispersed Nuclear Elements).

DNA spaziatore e di collegamento posto tra i geni e le altre regioni di DNA senza una definita funzione, né definito carattere di sequenza consenso simile in tutti i DNA spaziatori, può includere SINES E LINES (dimensione media 75kb).

>-------------<

DNA codificante proteine,

In questo DNA è incluso il DNA delle regioni regolatrici, degli esoni ed introni dei geni umani (dimensione media di un gene è 27kb, che varia da 1,4kb a 2400kb). La densità media dei geni nei cromosomi è 1 gene ogni 100kb (27kb il gene media + 75kb il DNA spaziatore medio).
I geni possono esistere come geni in copia singola, geni duplicati identici o parzialmente modificati (es. geni delle globine alfa, beta, gamma e delta) e geni non funzionanti: pseudogeni. Gli pseudogeni hanno sequenza simile e locus vicino al gene funzionante da cui sono originati mediante duplicazione, hanno mutazioni, accumulate dopo la duplicazione, che li rendono inattivi a codificare proteine funzionanti.
Si è stabilito mediante calcoli che i geni umani siano circa 30.000 e che codifichino circa 500.000 proteine.

La parte codificante di tutti i geni (i soli esoni) è circa il 2% del DNA genomico totale, mentre tutti i geni, completi delle parti funzionali al 5' ed al 3' e degli esoni e degli introni, è circa il 30% del DNA genomico. Il rimanente 70% è DNA spaziatore e soprattutto DNA di sequenze ripetute. Data l'alta variabilità delle dimensioni dei geni, i valori percentuali sono molto approssimativi perché calcolati utilizzando la dimensione media ed il numero medio dei geni. Per calcolare il numero totale dei geni di un genoma si contano uno o più tipi di sequenza nucleotidica ritenuti essere comuni a tutti i geni (es. sequenze segnale di splicing).

Tabella 1-1. Dimensioni dei geni umani e dei loro prodotti*

La dimensione media dei geni è	27kb (varia da 1400b a 2400kb)
Il numero medio di esoni è	9 (varia da 1 a 363)
La dimensione media degli esoni (esclusi gli esoni al 3' che in genere sono più lunghi) è	122b (varia da meno di 10b a 7.600kb)
La dimensione media degli introni non è calcolabile data l'enorme variabilità di dimensione degli introni, dimensione che è in relazione diretta alle dimensioni del gene (più grande è il gene più grandi sono i suoi esoni). Gli introni possono includere sequenze SINES e LINES.	
La dimensione degli introni varia da 10b a 8.000b.	
La dimensione media degli mRNA è	2.600b (varia da circa 900b a 115.000b).
La dimensione media delle proteine è	500-550 aminoacidi (varia da poche decine a 38.138 aminoacidi)

* da Strachan T. and Read A.P. (2004) Human Molecular Genetics. 3rd ed., Bios, UK, parzialmente modificato.

La capacità di un singolo gene di codificare più di una proteina è data dallo splicing alternativo del suo pre-mRNA e dalle diverse modificazioni covalenti post-traduzionali (es. glicosilazione, fosforilazione, acetilazione, metilazione, perdita di un peptide) che può avere la proteina da esso codificata. Le modificazioni covalenti devono essere stabili e la proteina attiva deve esistere in almeno due forme, es. non modificata e modificata-fosforilata. La modificazione covalente in genere modifica la funzione o è associata ad una localizzazione subcellulare.
Sono considerate prodotto di un unico gene le proteine modificate reversibilmente come nella fosforilazione che regola l'attività di una proteina. Egualmente le proteine prodotte dai diversi alleli di uno stesso gene sono considerate come un unico prodotto genico, sebbene possano avere proprietà sottili/subdole diverse (appendice D).

DNA codificante con sequenze ripetute in tandem: alcuni istoni, rRNA, 5sRNA e tRNA.

In biologia molecolare "In tandem" ha il significato inglese di "posto uno dopo l'altro" con due o più ripetizioni.

Nell'uomo i geni degli istoni (proteine nucleari) sono ripetuti e molti di essi si trovano in due gruppi (cluster) sul cromosoma 6. Gli istoni, sintetizzati in fase S del ciclo cellulare (fase di sintesi del DNA) sono senza introni e sono espressi

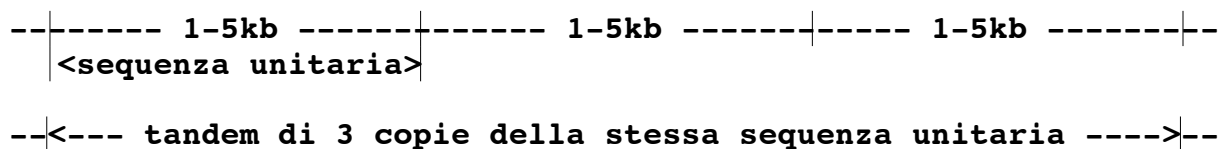
da geni ripetuti, mentre gli istoni sintetizzati nelle cellule quiescenti sono codificati da geni in copia singola che contengono introni.
I geni che codificano il pre-rRNA sono circa 250, i 5sRNA sono codificati da circa 2000 geni ed ogni singola specie di tRNA è ripetuta 10-100 volte. In totale i geni delle diverse specie molecolari di tRNA sono circa 1300.

DNA non codificante, ripetitivo e disperso nel genoma

DNA di sequenze ripetute in tandem.

E' costituito da sequenze con un numero variabile di ripetizioni in tandem, localizzate su più loci. VNTR (Variable Number of Tandem Repeats) sequences, dette anche Simple Tandem-Ripetitive Sequences (STRS).
Le sequenze VNTR sono costituite da sequenze unitarie ripetute in tandem, ed i tandem sono sparsi in loci diversi del genoma di tutti i cromosomi della specie umana ed anche riuniti in gruppi in particolari regioni del DNA (es. intorno al centromero, nei telomeri).

```
--|------ 1-5kb -----|------ 1-5kb ------|----- 1-5kb -------|--
  |<sequenza unitaria>|

--|<--- tandem di 3 copie della stessa sequenza unitaria ---->|--
```

In Drosofila, il DNA delle sequenze VNTR si trova principalmente intorno al centromero e vicino ai telomeri, mediante centrifugazione si separa facilmente dal rimanente DNA perché forma frammenti di DNA lunghi anche 10^5b. Per questo inizialmente fu chiamato DNA satellite del più abbondante DNA non caratterizzato da queste sequenze. Il termine satellite è ancora in uso, anche se non tutto il DNA delle sequenze VNTR può essere separato nello stesso modo.

Le sequenze VNTR hanno la sequenza unitaria conservata e ripetuta in tandem. Una data sequenza VNTR può avere più loci (posizioni subcromosomiche) nel genoma umano, e la posizione dei suoi loci è identica nei cromosomi di ogni individuo della specie umana. La stessa sequenza può avere un numero uguale o diverso di ripetizioni in tandem in loci diversi del genoma di uno stesso individuo e nello stesso locus nel genoma di individui diversi.
Data la diploidia uno stesso individuo può avere nello stesso locus dei due alleli una sequenza VNTR con un numero diverso di ripetizioni in tandem (figure 3-15 e 3-16).
Una sequenza VNTR che risulti avere un numero di sequenze in tandem diverso nello stesso locus di individui diversi è detta polimorfica per quel dato locus. Il polimorfismo rende la sequenza VNTR utilizzabile come marcatore genetico.
La sequenza unitaria di una sequenza VNTR può differire da un'altra per una o più basi costituendo una sequenza VNTR distinta dall'altra.
La lunghezza delle sequenze unitarie delle varie specie delle sequenze VNTR è molto diversa da 1 a circa 5kb ed anche la ripetizione in tandem varia in genere

da 1 a migliaia di volte per costituire (in un dato locus) una schiera di ripetizioni costituita da 100b - 20.000b come è nei minisatelliti (Tabella 1-2).
I loci, in cui è localizzata una data sequenza VNTR, possono essere da pochi ad oltre 50.000 come per i microsatelliti. Pertanto le sequenze VNTR possono variare tra loro per una diversa sequenza unitaria, per un diverso numero di ripetizioni in tandem e per un diverso locus.
Nell'uomo le sequenze VNTR sono classificate in base alle dimensioni della sequenza unitaria ed alla localizzazione cromosomica (Tabella 1-2).
Il DNA minisatellite ipervariabile è una famiglia di sequenze VNTR. Le diverse sequenze unitarie minisatellite variano molto in dimensione e molto meno in sequenza perché hanno come "sequenza base" (core sequence) una sequenza di 10b (GGGCAGGAXG) o di 16b (GGAGGTGGGCAGGAXG) (X può essere qualsiasi nucleotide). Ripetizioni diverse delle sequenze base costituiscono la sequenza unitaria dei diversi minisatellite. La variabilità è anche molto alta nel numero di ripetizioni e quindi nella dimensione dei tandem che variano da 100 a 20.000b distribuiti su 1-1000 loci ed in genere con percentuali di eterozigosi vicine al 100%. I loci sono dispersi nel genoma per lo più vicino ai telomeri degli autosomi.
Il DNA microsatellite, chiamato anche "Ripetizioni di Sequenze Semplici" (SSR, Simple Sequence Repeats"), é costituito da sequenze ripetute con sequenza unitaria di 1-4 basi, le ripetizioni in tandem sono relativamente poche (4-40) per cui i tandem in genere non superano 100b. Si calcola che le sequenze dei microsatelliti costituiscono circa il 2% del genoma (circa 60Mb) e che i loro loci, dispersi in tutto il genoma, siano oltre 180.000. Le sequenze unitarie dei microsatelliti più comuni sono 10. Si ritiene che il polimorfismo dei microsatelliti sia generato soprattutto dallo slittamento dei filamenti di DNAss durante la replicazione del DNA (figura 3-17).
I DNA microsatelliti costituiti dalla ripetizione di una base sono per lo più di (A)n/(T)n; sequenze poli-A sono poste nelle sequenze ripetute Alu (vedere dopo SINES) e complessivamente costituiscono lo 0,3% del genoma (circa 9Mb). Le ripetizioni di G/C sono rare. I DNA microsatelliti, costituiti dalla ripetizione di due basi, più comuni sono CA/TG, costituiscono complessivamente lo 0,5% del genoma (circa 15Mb), sono altamente polimorfici e dispersi nel genoma su oltre 83.000 loci (uno ogni 36.000b). I microsatelliti AT/AT sono dispersi su circa 60.000 loci (uno ogni 50.000 basi) ed i microsatelliti CG/CG sono su 24.000 loci (uno ogni 125.000b). Notare che le coppie di basi e quelle a loro complementari sono scritte ambedue in direzione 5'-3' (es. CA/TG)
Le ripetizioni CG/GC sono molto rare, come relativamente rare sono le sequenze VNTR di 3 basi (circa 5.000 loci/genoma) e quelle di 4 basi.
I microsatelliti costituiti da ripetizioni di 2-3 basi, altamente polimorfici e stabili nelle generazioni umane sono utilizzati come marcatori di loci di geni e per l'identificazione genetica degli individui.

Tabella 1-2. Sequenze umane VNTR

Sequenze VNTR	Sequenza unitaria numero di basi	Ripetizioni in tandem	Numero di loci	Posizione cromosomica
Megasatelliti	3-5kb	20-40	50-400	X, 4q, 19q
Satelliti	5-171b	500-500.000	---	centromerico in tutti i cromosomi
Minisatelliti altamente polimorfici	6-64b	10-2000	1-1000	telomerico e disperso in tutti i cromosomi
Microsatelliti altamente polimorfici	1-4b	4-40	oltre 180.000	disperso in tutti i cromosomi

Particolarità ed insidie delle sequenze ripetute in tandem.

La ripetizione in tandem di un tratto di sequenza identica (es. triplette) di DNA è causata da slittamenti di un filamento di DNAss (intracromatidici) durante la replicazione del DNA (figura 3-17) o ricombinazione ineguale (figura 3-18). Gli stessi meccanismi, se operanti in senso inverso, possono portare ad una riduzione delle ripetizioni in tandem. Quando questi processi avvengono durante la meiosi la variazione della ripetizione in tandem è trasmessa alla generazione successiva generando così una nuova variante di quella data sequenza VNTR. I polimorfismi VNTR attuali sono il risultato delle varie aggiunte/perdite di sequenze unitarie che casualmente sono avvenute in individui diversi durante i lunghi tempi della nostra evoluzione. Si assume che il polimorfismo dei minisatelliti si sia formato principalmente per ricombinazione ineguale e quello dei microsatelliti per slittamenti intracromatidici durante la replicazione del DNA.

La funzione dei microsatelliti è ignota. Ripetizioni in tandem CT/TG possono adottare la conformazione Z-DNA in vitro, ma non è dimostrato che lo facciano in vivo.

In una sequenza nucleotidica la citosina che ha al suo 3' una guanina (CpG) tende a subire una metilazione con successiva deaminazione che la converte in timina (TpG) (appendice B). Si ritiene che questa mutazione spontanea sia responsabile della scarsa presenza di VNTR CG/GC che contrasta con la grande

presenza di VNTR CT/AG (vedere sopra). La frequenza di CpG nel genoma è il 20% meno di quella che ci si aspetterebbe se non ci fosse stata durante l'evoluzione una continua conversione di CpG in TpG (per mutazione spontanea) e nel filamento opposto CpA (dopo duplicazione del DNA o riparazione, appendice C). Questo è un esempio di come il meccanismo genetico umano sia magnificamente evoluto sebbene costretto a rispettare la natura chimica delle basi, che non sono oggetti chimicamente inerti.
Tuttavia esistono "isole-CpG": sequenze in cui CpG sono presenti con frequenze maggiori del 50%. Isole-CpG si trovano nella regione <promotrice-primo esone> di geni espressi in più tessuti (come geni house keeping), mentre si trovano distanti ed a valle dal 3' del sito di inizio della trascrizione nel 40% dei geni espressi in uno/pochi tessuti. Le citosine delle isole CpG in genere non sono metilate mentre in genere lo sono le CpG sparse nel genoma.
I microsatelliti sono in genere localizzati in regioni non codificanti: intergeniche e introni. Alcune sequenze VNTR con sequenza unitaria di tre basi si trovano nelle zone promotrici e codificanti dei geni e la loro espansione in tandem durante la replicazione del DNA porta ad alterare l'espressione del gene o la proteina codificata causando stati patologici gravi (vedere appendice E).

<u>Sequenze singole localizzate su più loci</u>.
Queste sequenze sono anche dette "Intermediate repeat DNA".
Si ritiene che esse siano originate da mRNA cellulari (circa 2kb) o RNA virali (5-7kb) che sono stati convertiti in DNA ed inseriti nel DNA della cellula. Esse sono originate da particolari trasposoni (elementi genetici mobili) detti retrotrasposoni o retroposoni, ormai non più attivi (fossili) dopo una lunga evoluzione e colonizzazione del genoma umano. I retroposoni sono segmenti di DNA che codificano una trascrittasi inversa e si trasferiscono nel DNA cromosomico mediante un intermedio di RNA. L'RNA è convertito in DNA dalla trascrittasi inversa e poi è inserito in loci diversi dei cromosomi e in questo modo diviene ripetitivo e disperso. Alcune di queste sequenze hanno ai loro estremi sequenze ripetute simili (stesso orientamento) che si ritiene abbiano la funzione di integrare il trasposone nella sequenza nel DNA cromosomico.
L'attribuzione dell'origine da mRNA è basata sul fatto che queste sequenze hanno caratteristiche di geni ma non hanno introni e talvolta hanno segmenti di poli-A (sono pseudogeni originati da mRNA). L'origine virale di alcune sequenze è dedotta dalla caratteristica virale delle stesse sequenze. Trasposizioni con un intermedio di DNA (ricopiato dal DNA genomico) sono rare.
La sequenza unitaria è in genere una sequenza consenso perché spesso le sequenze di una stessa famiglia sono molto simili ma non identiche. Nell'uomo (in base alle dimensioni della sequenza consenso) le sequenze disperse isolate nel genoma sono divise in: SINES e LINES.
<u>SINES</u> (Short Interdispersed Nuclear Elements) nei mammiferi, uomo incluso, sono sequenze di 130-1300b con circa l'80% di identità tra individui della stessa specie (50-60% tra individui di specie diverse) e ciò è interpretato come indice di un progenitore comune (una sequenza simile a quella di

mammifero è stata trovata anche in E. coli). La più abbondante SINES nell'uomo (ed in altri mammiferi) è la sequenza Alu perché molte, ma non tutte le sue sequenze, includono un sito di riconoscimento per l'enzima di restrizione AluI. La sequenza consenso della famiglia Alu è circa 280b, ripetuta con circa $1x10^6$ copie (complessivamente 7-9% delle basi del genoma), dispersa nel genoma. Alu è simile ad un particolare oligoRNA di 294 nucleotidi (RNA 7SL). Questo RNA fa parte di un complesso ribonucleoproteico (RNA particle) coinvolto nella secrezione delle proteine attraverso la membrana del reticolo endoplasmatico. Alcune sequenze Alu sono trascritte in RNA breve e di breve vita, la funzione del quale è ignota. Le sequenze Alu sono disperse nel genoma anche all'interno di geni (introni e regioni non codificanti al 5' ed 3' del gene), tuttavia senza avere posizioni specifiche e mai all'interno delle parti codificanti (esoni). Le sequenze Alu hanno ai loro estremi due sequenze (una per estremo) dirette (stesso orientamento) con probabile funzione di riconoscimento per il loro movimento/inserimento nel DNA.

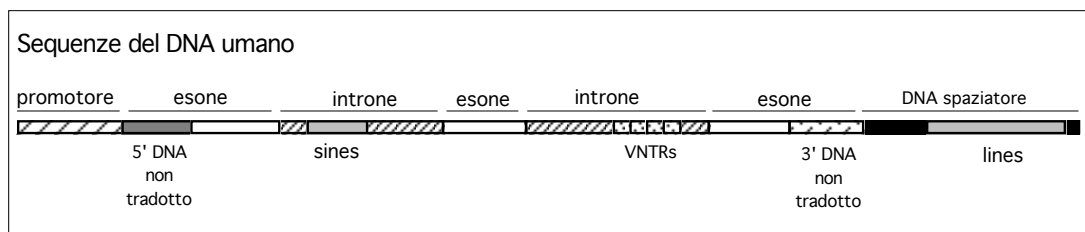
LINES (Long Interdispersed Nuclear Sequences), sono sequenze simili alle sequenze SINES ma più lunghe. Nei mammiferi (a differenza di altre classi di animali) si ipotizza che siano formate da retroposoni originati da RNA cellulari, piuttosto che da RNA virali. Nell'uomo e nel topo (Mus musculus) la sequenza consenso può essere di 1000-6000b. La famiglia più abbondante è la sequenza LINE-1 che, come Alu, ha agli estremi sequenze dirette. Nell'uomo, LINE-1 ha una sequenza consenso di 6.100b (non tutte le LINE-1 sono così lunghe: alcune solo 1400b) ed è ripetuta $6x10^4$ - $1x10^5$ volte per un totale di circa 10^8b. Le sequenze LINE-1 hanno alcune caratteristiche dei retroposoni ed alcune di esse trascrivono RNA.

	Dimensioni dei contenitori di DNA	DNA Numero di coppie di basi	DNA PM	DNA Lunghezza	Rapporto dimensioni DNA/ contenitore	DNA Lunghezza $x5x10^6$
batteriofago Øx174	25nm	$5,4x10^3$	$3,3x10^6$	1,9μm	76	9,95cm
batteriofago λ	190nm	$4,9x10^4$	$3,1x10^7$	18μm	95	8,75cm
batteriofagi T2 e T4	210nm	$1,8x10^5$	$1,1x10^8$	0,6mm	2857	33,3mm
Escherichia coli	2μm	$4,7x10^6$	$2,9x10^9$	1,7mm	850	8,5m
uomo	3-10μm (nucleo)	$6x10^9$ (diploide)	$3,7x10^{12}$ (diploide)	2,2m (diploide)	220.000*	11.000Km
cromosoma umano	7μm (mitotico)	$2,4x10^8$	$1,5x10^{11}$	8,6cm	12.286	430Km
gene strutturale (medio)		10^3	$6,2x10^5$	0,36mm		1,8m
mitocondrio	0,5-1μm	$1,7x10^4$ una molecola	$1,1x10^7$	6mm	6**	30m
coppia di basi		1	618	0,36nm		1,8mm

Tabella 1-3. Questo schema ha lo scopo di mostrare le grandi dimensioni del DNA rispetto ai contenitori (capsula virale, cellula batterica, nucleo di una cellula eucariotica e mitocondrio) che lo contengono e quindi l'alto grado di impacchettamento del DNA. Tutto ciò è straordinario. La lunghezza del DNA contenuto nel cromosoma 1 umano (il più lungo) è più di 8cm, per cui se il DNA non venisse impacchettato per azione degli istoni ed altre proteine, uomini e donne avremmo una fine e fitta peluria fatta di filamenti di DNA emergenti per circa 8 centimetri dalle cellule della pelle. Allineando il DNA (circa 2m) contenuto in tutte le cellule (circa 10^{14} cellule) di un uomo si forma un filo di DNAds lungo ($2x10^{11}$ km) tanto da andare dalla terra al sole (circa $1,5x10^8$km) e tornare per 1300 volte oppure sufficiente ad avvolgerlo intorno alla terra (circa $4x10^4$km) per 5 milioni di volte. Sebbene questi valori siano approssimativi danno una chiara idea della grande quantità di DNA presente nel corpo umano. Nell'ultima colonna a destra, le lunghezze del DNA sono state moltiplicate per $5x10^6$ per avere misure con le quali si ha più pratica e quindi è più facile fare confronti. Con l'ingrandimento di $5x10^6$, una coppia di basi ha le dimensioni di circa 1,8mm nella direzione della lunghezza del DNA. 1,8 mm è circa la larghezza che hanno questi simboli delle basi: A, T, G e C. Scrivendo con le stesse dimensioni una dopo l'altra, tutte le coppie di basi del genoma umano diploide di una singola cellula si costruirebbe una striscia lunga 11.000km, circa la distanza il linea d'aria da Pisa a Madison, Wisconsin, USA. La scrittura dell'intero genoma aploide umano occuperebbe 3000 libri di 500 pagine pertanto la sequenza del genoma umano è conservata nelle banche dati ed è consultabile solo per via elettronica mediante specifici programmi di gestione.
* Il DNA aploide umano è diviso in 23 cromosomi.
** I mitocondri di mammifero contengono 5-10 molecole di DNA circolare di 16.569b.
(Dati parzialmente modificati da Lehninger A.L., Nelson D.L. and Cox M.M. (1994) Principi di Biochimica. Zanichelli, Bologna. Watson J.D., Gilman M., Witkowski J. and Zoller M. (1992) Recombinant DNA,2nd ed., Scientific American Books, Freeman, USA. Alberts B., Bray D., Lewis J., Raff M., Roberts K. and Watson J.D. (1991) Biologia Molecolare della Cellula, 2a ed, Zanichelli; Connor J.M. and Fergusson-Smith M. A. (1991) Essential Medical Genetics, 3rd ed. Blackwell, London).

Alcune caratteristiche del genoma umano

Il corredo cromosomico di ogni individuo della specie umana è costituito da 22 coppie di cromosomi omologhi (autosomi) e da due cromosomi sessuali (2 X o X e Y). In ogni individuo umano, la molecola di DNA dello stesso cromosoma include le sequenze degli stessi geni e delle stesse sequenze ripetute e non ripetute che sono poste nelle stesse posizioni subcromosomiche fisiche (stessi loci). I geni umani sono indicati con le stesse sigle e nomi, codificano le stesse proteine che hanno gli stessi nomi e sigle e nei vari individui hanno la stessa attività molecolare e la stessa funzione cellulare. Le differenze genetiche degli individui dello stesso sesso dipendono dal polimorfismo dei geni, cioè dall'esistenza di uno stesso gene in due o più forme alleliche aventi diversa sequenza anche per una singola base nelle sequenze che regolano l'espressione del gene e/o nelle sequenze codificanti proteine.
Il sesso nella specie umana dipende da due cromosomi sessuali X e Y ed è determinato al concepimento dagli spermi che corredati di un cromosoma X o Y portano alla formazione di uno zigote che può essere XX o XY, dato che le uova umane hanno solo il cromosoma X. Il cromosoma Y è trasmesso solo per via maschile e il cromosoma X nei maschi e un cromosoma X nelle femmine ha sempre origine materna. Il DNA mitocondriale è solo di origine materna. Gli individui maschi e femmine hanno mitocondri originati solamente da quelli presenti nell'uovo dato che i mitocondri degli spermi non vengono trasferiti nell'uovo al momento della fecondazione. Gli uomini e le donne hanno gli stessi cromosomi autosomici, lo stesso cromosoma X (quindi gli stessi geni di questi cromosomi). Il cromosoma X, aploide nell'uomo, è da considerare tale anche nella donna data l'inattivazione casuale di uno dei due cromosomi X negli embrioni femminili di 16 giorni. Ciò causa l'inattivazione della quasi totalità dei geni del cromosoma X in tutte le cellule somatiche della donna, che risulta essere un mosaico di cromosomi attivi materni e paterni. Uomo e donna differiscono per quasi tutti i geni contenuti nel cromosoma Y (il 95% del cromosoma Y non ricombina con il cromosoma X). Sul cromosoma Y è localizzato il gene SRY, responsabile della determinazione del sesso maschile ed altri 3-4 geni coinvolti nella spermatogenesi. Il gene SRY codifica un fattore di trascrizione che induce lo sviluppo del testicolo nell'embrione di 5 settimane, attivando geni disposti su altri cromosomi. Successivamente durante la vita dell'individuo, il testicolo produrrà ormoni per lo sviluppo dei caratteri sessuali secondari attivando geni disposti sugli altri cromosomi, cromosoma X incluso: 99 geni del cromosoma X codificano proteine espresse nei testicoli. Il cromosoma Y del canguro contiene solo il gene SRY (unica differenza genetica tra maschi e femmine) che induce il differenziamento maschile attivando geni che (ovviamente) sono presenti anche nelle femmine.
Pertanto, l'uomo differisce geneticamente dalla donna per 4-5 geni posti sul cromosoma Y, le uniche differenze di geni presenti tra gli individui di sesso diverso della popolazione umana.

Ogni individuo umano dello stesso sesso è geneticamente diverso dagli altri perché ha una combinazione di alleli degli stessi geni polimorfici (si assume che lo siano quasi tutti i geni) diversa rispetto a quella di ogni altro individuo. L'eterogeneità della composizione degli alleli interessa anche individui di sesso diverso ed interessa i geni polimorfici posti sugli autosomi e sul cromosoma X (appendice D).
Questa eterogeneità genetica esprime caratteristiche morfologiche e funzionali diverse (fenotipo diverso) nei diversi individui umani (diversamente alti, forti, veloci, intelligenti, ecc.). La responsabilità delle diversità di queste caratteristiche è attribuita principalmente alla eterogeneità degli alleli dei circa 1.000 geni regolatori di geni (vedere dopo in questo stesso paragrafo). L'altra responsabilità è attribuita ai circa 29.000 geni che codificano proteine che sono responsabili della morfologia e delle funzioni delle cellule e degli organi (enzimi, ormoni, recettori degli ormoni, proteine segnale, strutturali e proteine contrattili, ecc.). Le proteine codificate da alleli diversi di uno stesso gene (isoproteine), hanno tutte la stessa attività molecolare, lo stesso tipo di regolazione e la stessa funzione cellulare. Tuttavia esse possono possedere proprietà dette minori o subdole: valori diversi di concentrazione cellulare, di stabilità (vita più o meno lunga) o attività molecolare, diversa affinità per molecole endogene (metaboliti o molecole segnale), sensibilità diversa verso molecole esogene (farmaci, anestetici, allergeni, inquinanti, veleni). Le proprietà minori sono chiamate anche subdole quando si manifestano con effetti in genere negativi in particolari condizioni di alimentazione, di ambiente, di normali attività fisiologiche (es. digiuno, sforzo muscolare) o durante terapie. Si assume che le diverse combinazioni di proprietà minori e subdole contribuiscano alla costituzione morfologica e funzionalmente diversa degli individui e così anche alla formazione delle patologie complesse (capitolo 4 ed appendice E). Si assume anche che le molecole provenienti dagli alimenti o dall'ambiente influenzino la costituzione degli individui interagendo in maniera diversa in relazione alle proprietà minori e subdole delle proteine.
Si ritiene che la maggior parte degli alleli dei geni sia ben distribuita nella popolazione umana di una data area geografica, ad esempio dell'Italia, sulla base di una semplice considerazione. Gli ascendenti di ogni italiano vivente, risalendo la genealogia, aumentano secondo le potenze del 2: 2 genitori, 4 nonni, 8 trisavoli. Assumendo generazioni di 25 anni e risalendo 40 generazioni si arriva all'anno mille, pertanto un individuo attualmente vivente avrebbe complessivamente 2^{40} antenati, più di 1.000 miliardi di antenati. Considerando 80 generazioni si arriva all'anno zero, 2^{80} antenati, un numero superiore a 1,2 $x10^{24}$. I numeri sono sicuramente superiori a quelli della popolazione italiana di quei tempi, includendo anche le possibili immissioni da altri territori di altre popolazioni. Essendo originati da una popolazione più piccola, ogni italiano ha molti ascendenti in comune con ogni altro italiano e quindi porta i relativi alleli. La conseguenza di ciò è che siamo tutti imparentati e le differenze morfologiche e funzionali tra gli italiani dipendono da combinazioni di alleli diversi. Nello stesso

modo si spiega anche la vasta distribuzione degli alleli delle patologie poligeniche in Italia e nel mondo (capitolo 4 ed appendice E).
Con l'avvento delle tecnologie del DNA è stato possibile valutare il polimorfismo delle varianti modificanti il fenotipo normale (quantità e qualità delle proteine) e le varianti che non lo modificano (non è modificata la quantità e la qualità delle proteine). La genetica classica analizzava solo le varianti che causavano variazioni del fenotipo normale perché erano le uniche che potevano essere osservate non essendo ancora disponibili le tecnologie del DNA (capitolo 3).
Gli alleli mutati che causano patologie (patologie monogeniche) non sono considerati come facenti parte del polimorfismo, che per definizione include solo le varianti genetiche non-patologiche. Tuttavia i meccanismi molecolari che generano il polimorfismo normale e gli alleli patologici sono identici (errori di sintesi del DNA, ricombinazione genetica tra sequenze omologhe non identiche, effetti di radiazioni o di reagenti, integrazione nel genoma di DNA virale). E' solo questione di fortuna o sfortuna, quella casualità che James Watson chiama "i dadi genetici".
Nella sequenza del DNA dei cromosomi omologhi possono esserci differenze di lunghezza, relativamente piccole, dovute al polimorfismo delle sequenze ripetute VNTR (più meno lunghe), dagli inserimenti di DNA di trasposoni o di retroposoni avvenuti durante i lunghi tempi dell'evoluzione. Queste differenze di lunghezza del DNA cromosomico non alterano l'attività molecolare del DNA genomico perché essa non è basata sull'esatta distanza tra sequenze da uno stesso punto fisso di riferimento (es. sequenza del centromero), ma sul riconoscimento molecolare (appendice C) operato tra sequenze di DNAss complementari (es. la ricombinazione genetica omologa alla meiosi, appendice D), tra sequenze di DNAds e proteine (es. l'associazione dei fattori di trascrizione alle sequenze promotrici, appendice B).
Il meccanismo con il quale si appaiano i cromosomi alla meiosi è sconosciuto, tuttavia le relativamente piccole differenze di lunghezza del DNA cromosomico non impediscono la corretta formazione di chiasmi, singoli e multipli, tra cromosomi omologhi perché la loro formazione è sequenza specifica e avviene tra sequenze di DNA identiche dei due cromosomi omologhi. Le due anse del DNA tra due chiasmi possono avere piccole differenze di lunghezza di DNA per la presenza di sequenze diversamente ripetute, differenze difficilmente visibili al microscopio data la grande dimensione e l'alta condensazione del DNA dei cromosomi che si appaiano alla mitosi (10.000 volte rispetto al DNA in doppia elica, tabella 1-3 ed appendice B). Talvolta nella regione di DNA in ricombinazione (giunzione eteroduplice) le sequenze dei due alleli differiscono per una/poche basi e ciò può portare alla riconversione genica con perdita di eterozigosi (appendice D).
Nella meiosi, la formazione dei chiasmi e la distribuzione casuale dei cromosomi omologhi, garantiscono la formazione di un numero pressoché illimitato di gameti geneticamente diversi. La dimensione ed il numero dei cromosomi sono caratteristiche di ogni specie, esse sono fondamentali per l'avvenire di una corretta meiosi ed in particolare l'appaiamento e la segregazione dei cromosomi

omologhi al fine di produrre gameti aploidi per generare una prole diploide di soli cromosomi omologhi, quindi normale e feconda.
Durante gametogenesi, alla meiosi e nel giovane embrione si può avere la formazione di aberrazioni cromosomiche, cioè alterazioni nel numero (aneuploidia e poliploidia) o nella struttura dei cromosomi dei gameti e quindi negli organismi concepiti. Durante la meiosi prima della anafase-I, i chiasmi si comportano come centromeri tenendo uniti i cromosomi omologhi materni e paterni. Se dopo l'associazione, non avviene anche la ricombinazione, non si ha la segregazione dei cromosomi bivalenti e ciò causa la formazione di gameti con un numero non corretto di cromosomi. Le alterazioni strutturali dei cromosomi sono conseguenti a rotture di cromosomi. Quando un cromosoma si spezza, gli estremi spezzati tendono a riassociarsi ed i meccanismi di riparazione rilegano rapidamente e correttamente le parti di DNA spezzate; tuttavia quando si ha più di una frattura si può avere la riassociazione e legatura non corretta delle parti spezzate, ad esempio appartenenti a cromosomi diversi, perché i sistemi di riparazione non le distinguono. Le alterazioni strutturali includono la traslocazione (scambio di parti di cromosoma tra cromosomi diversi), la delezione (perdita di parte del cromosoma), la duplicazione di parti di cromosoma ed altri tipi di aberrazione che portano alla formazione di gameti con corredo cromosomico alterato.
Le alterazioni cromosomiche sono comuni, sono il 7,5% dei concepimenti e la frequenza di alcune aumenta con l'età della madre (aneuploidia) o con l'esposizione a radiazioni (traslocazioni). Tuttavia solo lo 0,6% dei nati vivi ha aberrazioni cromosomiche perché molte di esse (92%) causano aborti spontanei precoci e tardivi o nati morti. I portatori di aberrazioni cromosomiche, clinicamente sani (portatori di aberrazioni cromosomiche compensate dalla diploidia), hanno un alto rischio di generare portatori di aberrazioni non compensate. I portatori di aberrazioni non compensate in genere non hanno prole. Quindi anche se c'è sempre la formazione spontanea di aberrazioni cromosomiche, c'è anche una tendenza ad eliminarle prima della nascita o nelle generazioni successive con la dura legge della selezione naturale. Aberrazioni di numero e di struttura dei cromosomi avvengono anche se con minore frequenza anche durante la mitosi. Sono chiamate mutazioni somatiche e possono essere responsabili di individui mosaico e di patologie (vedere figure D-2 e E-2).
Il DNA genomico umano durante la meiosi va incontro a circa 30 eventi di ricombinazione per corredo aploide, uno o più eventi per cromosoma in relazione alla dimensione del cromosoma (capitolo 3). La frequenza di ricombinazione è legata al sesso: nella donna si hanno circa 70 chiasmi per meiosi e nell'uomo 55. Regioni di DNA di uno stesso cromosoma hanno frequenze di ricombinazione molto diverse ed esistono regioni anche includenti più di un gene che non sono andate incontro a ricombinazione per molti secoli (aplotipi). Gli aplotipi possono includere combinazioni di alleli normali e patologici caratteristici di una popolazione (alleli etnici). Gli alleli etnici sono caratteristici di popolazioni separate da altre fisicamente, culturalmente o per credo religioso, per cui gli

individui di queste popolazioni tendono a sposarsi tra loro ed un allele che subisce una mutazione nel tempo e si distribuisce nella popolazione rimane confinato in essa caratterizzandola. Gli alleli di un aplotipo sono trasmessi associati dai genitori ai figli per molte generazioni e con essi è possibile costruire specifici alberi genealogici (capitolo 3).
Durante la gametogenesi, ed in generale in ogni evento di replicazione del DNA (meiosi e mitosi) il corretto mantenimento del numero ed integrità dei cromosomi, delle sequenze del DNA e dei loro loci dipende dalla corretta replicazione del DNA che essendo lineare mantiene l'ordine originale della sequenza del DNA, dai corretti eventi di ricombinazione omologa (chiasmi) durante la meiosi, dall'efficienza dei meccanismi di riparazione del DNA e di sorveglianza dell'integrità del genoma.
Errori nella sintesi del DNA portano a variazioni della sequenza del DNA che possono essere patologiche dominanti o recessive, possono instaurare proprietà minori o subdole nelle proteine codificate o non causare alterazioni nelle proteine e quindi nel fenotipo. Le alterazioni cromosomiche, data la grande dimensione del DNA coinvolto, in genere hanno effetti così drammatici da causare la morte di chi le porta o quella dei figli che genera.
L'importanza del numero e dimensione dei cromosomi è dimostrato dalla presenza in natura di due specie di piccolo cervo, morfologicamente molto simili, aventi corredi cromosomici diploidi molto diversi. Il muntjac cinese ha 46 cromosomi ed il muntjac indiano ha 6 cromosomi nella femmina e 7 nel maschio. Le due specie sono fenotipicamente molto simili, si possono accoppiare ed avere prole che tuttavia è sterile. Si assume che le due specie siano geneticamente (numero e sequenze di geni) identiche ma che la loro speciazione sia stata causata dall'evoluzione dei cromosomi e non da quella dei geni. L'antenato comune aveva 46 cromosomi e la fusione degli stessi li avrebbe ridotti a 6-7 creando ad una nuova specie il (muntjac indiano), oppure l'antenato aveva 6-7 cromosomi che si sono frantumati in 46 dando vita al muntjac cinese.
Il particolare corredo cromosomico (23+3 femmine o 23+4 maschi) della prole generata dall'accoppiamento delle due specie di muntjac impedisce di operare una meiosi corretta e quindi di avere una prole feconda. L'appaiamento corretto dei cromosomi omologhi e la successiva distribuzione casuale dei cromosomi nei gameti appaiono come eventi impossibili, perché sarebbe necessario distribuire in un gamete i 23 cromosomi provenienti dal genitore, muntjac cinese, e nell'altro gamete i 3 o 4 cromosomi, ad essi omologhi, provenienti dall'altro genitore, il muntjac indiano. Oppure, ancora più difficile, produrre una prole normale con gameti aventi un corredo cromosomico incompleto che si compensano reciprocamente (il cromosoma che manca in un gamete di un genitore deve essere presente nel gamete dell'altro genitore).
Anche se il caso di speciazione del muntjac è forse unico nei vertebrati, esso dimostra che non solo il <u>complesso (numero e specie molecolari) dei geni</u> specifica una data specie ma anche il <u>numero e la dimensione dei cromosomi</u> sui quali sono distribuiti gli stessi geni (cioè numero, dimensione e disposizione dei

loci dei geni) sono specie specifici. Ciò è imposto dal meccanismo della meiosi necessario per la ricombinazione del genoma e per produrre gameti aploidi al fine di mantenere una corretta diploidia nelle generazioni.
I recenti studi di genomica comparata (confronto del genoma totale di più specie) hanno migliorato la conoscenza dei possibili meccanismi genetico molecolari dell'evoluzione delle specie e del differenziamento degli organi[48].
Un importante meccanismo di evoluzione opera mediante la duplicazione dei geni e con le successive modificazioni (mutazioni) delle sequenze regolatrici e codificanti del gene duplicato.
Un esempio è dato dall'emoglobina umana, tetramero (alfa-beta)$_2$ dove la duplicazione del gene codificante la catena-beta (espressa dopo la nascita e negli adulti) ha portato alla formazione della catena-gamma. La catena-gamma è espressa solo nel feto ed ha una sequenza diversa dalla globina-beta, pertanto l'emoglobina fetale (alfa-gamma)$_2$ ha un grado di affinità per l'ossigeno molecolare maggiore di quello della emoglobina adulta (alfa-beta)$_2$. La maggiore affinità per l'O_2 dell'emoglobina fetale favorisce il flusso del l'O_2 dall'emoglobina materna a quella fetale.
Quando il meccanismo della duplicazione e successiva mutazione interessa un gene regolatore di geni (esempio: il gene che codifica un fattore di trascrizione) si può avere la modificazione dell'espressione quantitativa e qualitativa di più geni ed anche l'espressione degli stessi geni in cellule di organi diversi da quelli originali. La sostituzione di una aminoacido nelle proteine che non regolano geni ha conseguenze limitate alla proteina stessa (proprietà minori o subdole) mentre la sostituzione di un aminoacido nelle proteine regolatrici di geni può avere conseguenze maggiori perché viene interessata l'espressione di altri geni. Questi cambiamenti di espressione di più geni possono far sorgere nuove e complesse capacità funzionali negli organi già presenti nell'individuo e renderlo più o meno atto alla selezione naturale (vedere dopo).
Si assume che i geni regolatori di geni siano circa 1000 e circa 100 di essi siano geni determinanti il modello di espressione di altri geni (pattern determining genes).
Uno di questi geni (Pax6) controlla lo sviluppo degli occhi in quasi tutti gli animali e si ritiene che cambiamenti del modello di espressione di Pax6 siano responsabili della diversità morfologica degli occhi di animali diversi. Questo gene è normalmente espresso negli occhi durante il loro sviluppo, quando è espresso in altri tessuti determina la formazione di occhi anche in quei tessuti (occhi ectopici), ad esempio sulle gambe o ali della drosofila. Si ipotizza che la diversa posizione degli occhi nelle diverse specie animali dipenda dal cambiamento del modello di espressione di Pax6 durante l'evoluzione. Nelle chiocciole, la particolare localizzazione delle macchie oculari all'estremo dei peduncoli, è stata messa in relazione ad una alterazione del modello di espressione di Pax6 che in molti animali colloca gli occhi in affossamenti della testa in posizioni diverse (esempio: frontali o laterali). La transfezione in drosofila del gene Pax6 di seppia, modificato nel suo modello di espressione, porta alla formazione di occhi ectopici sulle gambe o ali, come accade in

drosofila quando il suo gene Pax6 ha il modello di espressione alterato. Le proteina Pax6 di drosofila e di seppia hanno solo il 30% di identità di sequenza, tuttavia hanno attività simili: ambedue le proteine determinano la formazione di occhi ectopici nelle gambe ed ali delle drosofile transgeniche. Queste osservazioni hanno suggerito che, durante l'evoluzione per la creazione di occhi morfologicamente diversi, siano stati più importati i cambiamenti di espressione del gene Pax6 che i cambiamenti di sequenza e quindi di attività molecolare della proteina Pax6.
Watson[48] ipotizza che cambiamenti di morfologia possono risultare anche da cambiamenti della sequenza e quindi della funzione della proteina codificata dai geni determinanti il modello di espressione. Durante l'evoluzione, un gene determinante il modello di espressione può subire un evento di duplicazione e le due proteine codificate, inizialmente identiche, possono divergere (modificare la sequenza durante l'evoluzione) ed acquistare proprietà di regolazione diverse: agire sull'espressione di gruppi di geni diversi in numero e/o qualità oppure passare da attivatrice ad inibitrice dell'espressione degli stessi geni. La sostituzione di uno o qualche aminoacido modifica l'affinità di associazione della proteina verso le sequenze di regolazione dei geni, aumentando o riducendo il numero dei geni bersaglio attivati o inibiti. Nell'evoluzione il cambiamento di funzione della proteina ha come conseguenza il cambiamento del modello di espressione dei geni bersaglio della proteina.
Un terzo meccanismo, che nell'evoluzione può aver causato diversità morfologiche, è attribuito a cambiamenti nella sequenza del DNA dei promotori o enhancer che sono regolati dai geni determinanti il modello. In questo caso il gene determinante il modello non ha subito modifiche nella sua espressione e la proteina codificata non ha cambiato attività. La sostituzione di una singola base su 200 può essere sufficiente a modificare molto l'affinità verso la proteina regolatrice e risultare in una diversa espressione dei geni bersaglio.
Enhancer con bassa affinità sono attivati solo da alte concentrazioni di proteina regolatrice mentre quelli ad alta affinità sono regolati anche da basse concentrazioni della proteina regolatrice. Regolando la concentrazione della proteina regolatrice si opera la regolazione differenziale dei geni aventi enhancer con valori diversi di affinità verso la stessa proteina. Si ipotizza che nell'evoluzione la duplicazione di un enhancer seguita da una mutazione che ne faccia variare l'affinità verso la proteina che lo regola o lo renda affine ad altre proteine regolatrici di geni, abbia portato a cambiamenti nel modello di espressione dei geni e a cambiamenti nella morfologia degli animali.
Il numero degli enhancer dei geni è molto importante al fine di poter conferire più di un modello di espressione di un gruppo di geni. A parità di numero di geni (esempio circa 20.000), la specie che ha una media di enhancer per gene superiore (esempio 3-4), ha un numero di modelli di espressione superiore (circa 50.000) contro i 30.000 modelli di espressione di una specie in cui circa 20.000 geni abbiano mediamente 1-2 enhancer.

Sommario:
Le strategie che durante l'evoluzione possono aver modificato l'espressione dei geni determinanti il modello e quindi la morfologia degli animali sono almeno tre:
1. Un gene determinante il modello può modificare il suo modello di espressione e di conseguenza modifica anche quello dei geni da esso controllati (geni bersaglio).
2. La proteina codificata da un gene determinante il modello subendo una mutazione può modificare la sua affinità per specifiche sequenze regolatrici di DNA e quindi attivare geni diversi e/o modificare la sua attività rovesciando i suoi effetti: da attivatrice a inibitrice (o viceversa).
3. Geni bersaglio possono modificare le loro sequenze di DNA regolatrici (promotori e enhancer) o acquisirne nuove (via duplicazione e mutazione) e quindi modificare il loro modello di espressione passando sotto il controllo di un differente gruppo di geni regolatori.

L'analisi comparativa dei genomi suggerisce che l'evoluzione dei vertebrati, ed in particolare quella dell'uomo, sia avvenuta principalmente mediante modifiche della sequenza dei geni determinanti il modello e/o della sequenza delle sequenze regolatrici dei loro geni bersaglio.
A questa conclusione portano vari dati sperimentali.
Il topo e l'uomo hanno circa lo stesso numero di geni (circa 30.000), 80% di questi geni hanno sequenze molto simili (più dell'80% delle stesse basi nella stessa posizione). Le proteine codificate da questi geni sono molto conservate ed hanno mediamente l'80% di identità di sequenza aminoacidica (stesso aminoacido nella stessa posizione).
Il rimanente 20% di geni umani e murini codificanti proteine sono derivati per duplicazione da quelli sopra indicati ed in relazione al numero degli eventi di duplicazione, il loro numero può essere diverso nelle due specie. Nel topo sono presenti più geni del citocromo P450 di quelli presenti nell'uomo, pertanto le isoforme umane del gene del citocromo P450 troveranno il corrispondente murino e faranno parte dell'80% di geni simili, mentre le rimanenti isoforme murine non trovando il corrispondente umano faranno parte del 20% di geni non simili a quelli umani. Esistono anche casi opposti in cui le copie umane di uno stesso gene sono maggiori di quelle murine, ma non esistono geni umani che siano completamente assenti nel topo. Inoltre esistono tratti dei cromosomi delle due specie che sono sintenici, cioè la sequenza dei loci dei geni nel DNA delle due specie è conservata.
La similarità tra uomo e topo è evidente anche nel fenotipo; con il topo abbiamo molte omologie: vie metaboliche, organi e sofisticate funzioni, forse John Steinbeck per altre vie l’aveva già capito.

Il grande incremento nel numero di geni osservato nei vertebrati (circa 30.000) rispetto agli invertebrati (circa 15.000) è dovuto principalmente alla duplicazione dei “vecchi” geni piuttosto che all’invenzione di nuovi. Ed anche questa osservazione favorisce l’ipotesi che la diversità morfologica degli animali dipenda da differenze di modello di espressione di proteine molto simili aventi la stessa attività molecolare. Analogamente il campanile, la cattedrale ed il

battistero del Campo dei miracoli di Pisa sono tre modelli costruiti con lo stesso tipo di pietra aventi poche dimensioni standard (come con il Lego).
Il confronto delle sequenze del genoma di scimpanzé con quello umano è ancora più strabiliante: mediamente il 2% di divergenza nella sequenza del DNA delle due specie, quando nelle ascidie, individui appartenenti a popolazioni diverse della stessa specie hanno differenze di sequenza del 2,5%. Inoltre la sintenia tra scimpanzé ed uomo è molto estesa per cui l'ordine e le distanze tra i geni sono altamente conservate. Gli alti livelli di conservazione del numero e delle sequenze dei geni dei genomi delle tre specie (topo, scimpanzé ed uomo), indicano che le differenze morfologiche tra le tre specie e così anche l'apparire improvviso di una nuova funzione durante l'evoluzione umana, dipendano da una differente espressione di geni aventi sequenze codificanti molto simili e non dalla presenza o dalla comparsa di nuovi geni in una specie (esempio l'uomo) che sono assenti in altre specie meno evolute.
Watson[48], per ipotizzare come un innovazione evolutiva possa essere sorta nell'uomo, propone di analizzare la parola (la capacità altamente precisa di comunicare con il linguaggio), una caratteristica che definisce molto bene gli esseri umani, mentre il linguaggio degli altri vertebrati, scimpanzé incluso, è rimasto molto rozzo.
La capacità di parlare dipende dalla precisa coordinazione dei piccoli muscoli della laringe e della bocca che è associata alla corretta espressione della proteina regolatrice FOXP2. La carenza di FOXP2 causa severi difetti nel parlare e gli individui affetti hanno difetti diversi nell'articolare le parole. Il gene FOXP2 è stato isolato da vari animali, uomo incluso e si è osservato che la proteina FOXP2 è altamente conservata: la proteina umana differisce per 2 e 3 aminoacidi rispettivamente da FOXP2 di scimpanzé e di topo. Due aminoacidi caratterizzano l'uomo: asparagina-303 e serina-325 (rispettivamente treonina e asparagina nello scimpanzé). Questi aminoacidi sono localizzati nel dominio che ha attività di repressore sull'espressione dei geni bersaglio della proteina FOXP2 e ciò porta ad ipotizzare[48] che la proteina umana sia incapace di inibire l'espressione di alcuni geni che determinano la buona coordinazione dei muscoli della laringe e della bocca e che invece la proteina FOXP2 di scimpanzé e di topo inibiscono.
Il cambiamento nella coordinazione dei muscoli della laringe e bocca durante l'evoluzione umana può essere stato causato da cambiamenti dell'espressione del gene FOXP2 o degli enhancer dei suoi geni bersaglio, cioè secondo i meccanismi 2 e 3 sopra indicati.
Watson[48] spiega che, sebbene queste considerazioni siano speculative, perché basate sull'analisi di solo tre genomi, esse possono dare l'indicazione di come poche mutazioni in pochi geni regolatori di geni possano aver causato il salto evolutivo dell'acquisizione di una caratteristica così importante come l'uso della parola. Ad esempio, un cambiamento (mutazione) nel DNA di una sequenza regolatrice del gene FOXP2 può conferire al gene un nuovo modello di espressione nell'encefalo umano che porta a sintetizzare la proteina FOXP2 in cellule di nuove ed appropriate regioni dell'encefalo e nel momento giusto del

loro sviluppo. I geni bersaglio della proteina FOXP2 poi possono codificare proteine che vengono espresse al momento giusto quando il bambino è più sensibile ad apprendere a parlare. Nelle bocche e faringi umane, l'espressione del gene FOXP2 sarebbe attivata nel tempo in cui il bambino è più suscettibile ad imparare a parlare attivando geni le cui proteine determinerebbero la corretta coordinazione dei muscoli oro-laringei.
Lo scimpanzé, sebbene abbia come l'uomo il gene FOXP2 che codifica una proteina pressoché identica, non sarebbe capace di un linguaggio articolato perché la proteina FOXP2 non sarebbe espressa nelle giuste aree cerebrali e/o nella bocca e nella laringe in tempi e quantità corrette.
La differenza di linguaggio tra uomo e scimpanzé sarebbe causata da piccole differenze (mediamente il 2%) nella sequenza del DNA di un gene determinante il modello di espressione di geni bersaglio che codificano le proteine che hanno la funzione di determinare la corretta coordinazione dei muscoli orolaringei. Mentre si assume che le piccole differenze (mediamente il 2%) di sequenza del DNA dei geni codificanti le proteine che nell'uomo e nello scimpanzé coordinano i muscoli oro-laringei, non possano essere, per le ragioni sopra indicate responsabili della diversa capacità di linguaggio delle due specie.
I dati ottenuti con la genomica comparata indicano che esistono almeno due gruppi di geni: i geni regolatori di geni che includono i geni determinanti i modelli di espressione (rispettivamente circa 1000 e 100 nei vertebrati) ed i geni codificanti proteine con altre attività (circa 29.000 nei vertebrati) che codificano proteine responsabili di funzioni cellulari e che hanno attività molecolari come l'associazione di leganti, la catalisi, la contrazione e il mantenimento di strutture, il trasporto di molecole, il metabolismo, la trasduzione e trasferimento dei segnali.
I diversi aspetti morfologici dei vertebrati di oggi dipenderebbero soprattutto dalla diversa espressione dei geni regolatori di geni, dalle differenze di sequenza delle proteine da essi codificate e dal numero e sequenza degli enhancer dei loro geni bersaglio piuttosto che dalle proteine responsabili delle funzioni cellulari.

Durante l'evoluzione dei primati, l'acquisizione da parte dell'uomo di funzioni anche molto complesse sarebbe stata causata da mutazioni puntiformi dei promotori o degli esoni dei geni regolatori di geni e/o degli enhancer dei loro geni bersaglio. All'acquisizione di nuove funzioni non parteciperebbero le proteine responsabili della fisiologia cellulare perché si è ripetutamente osservato, che nelle specie animali aventi storie evolutive molto diverse, le proteine omologhe possono avere differenze di sequenza grandi (fino oltre il 70% di aminoacidi sostituiti) tuttavia conservano la stessa attività molecolare e quindi anche la stessa funzione cellulare (capitolo 1, similarità delle proteine). Pertanto, la morfologia e l'attività funzionale di una cellula e di un organo sono potuti cambiare quando un nuovo modello di espressione ha portato a sintetizzare nelle cellule di un dato tessuto/organo una nuova combinazione di proteine funzionali.
I meccanismi proposti spiegano anche le differenze morfologiche e funzionali esistenti tra gli individui dell'attuale popolazione umana (appendice D).
Queste differenze sebbene piccole possono avere una grande importanza.
"Se il naso di Cleopatra fosse stato più corto, sarebbe cambiata l'intera faccia della terra." (Blaise Pascal).

Leggi e convenzioni della biologia molecolare

Nozioni minime per comprendere l'analisi del genoma umano
ovvero manuale del manovale genetico molecolare.

1. Basi molecolari della specificità dell'appaiamento dei due filamenti del DNA.
I due filamenti del DNA si appaiano specificamente e con alta stabilità in conseguenza dell'appaiamento di ogni singola coppia di basi che costituisce il DNA a doppio filamento (DNAds =Double Stranded DNA). L'appaiamento delle basi è sempre Adenina-Timina (A/T) e Guanina-Citosina (G/C) (regola di Chargaff) (figura 1-1). La specificità dell'appaiamento delle basi è determinata dalla formazione tra loro di legami ad idrogeno (due in A/T e tre in G/C) che avvengono per una precisa disposizione delle basi e in particolare dei loro gruppi in legame a H, ordinatamente per ogni coppia di basi che costituisce il doppio filamento. Si possono formare filamenti DNA-RNA ed RNA-RNA con la sola differenza che nell'RNA la timina è sostituita dall'uracile (U) e il desossiribosio dal ribosio.
E' sufficiente l'appaiamento di 20 basi per dare luogo ad un complesso di due filamenti molto stabile (circa 50kcal/mole a 37°C). Se il DNAds è ricco di coppie GC la stabilità del doppio filamento (a parità di lunghezza) è maggiore, dato che la coppia GC è stabilizzata da tre legami a idrogeno.

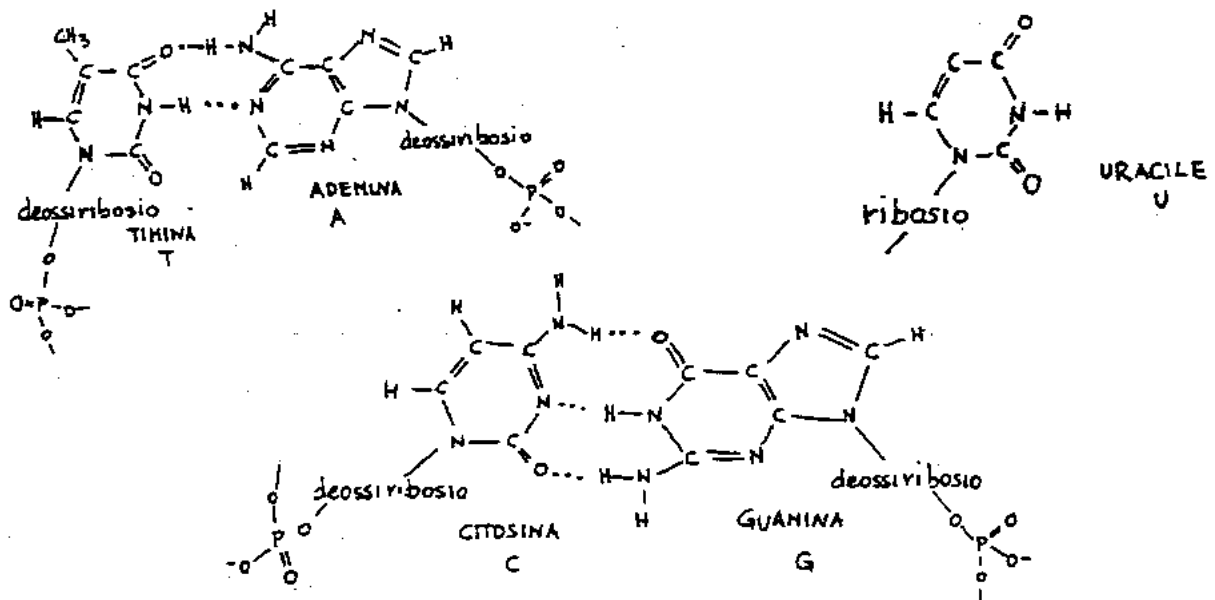



Figura 1-1. Legami ad idrogeno tra le coppie di basi Adenina-Timina (A/T) e Guanina-Citosina (G/C). L'Uracile (U) come la Timina (che è metil-uracile) si appaia con l'Adenina (A-U).

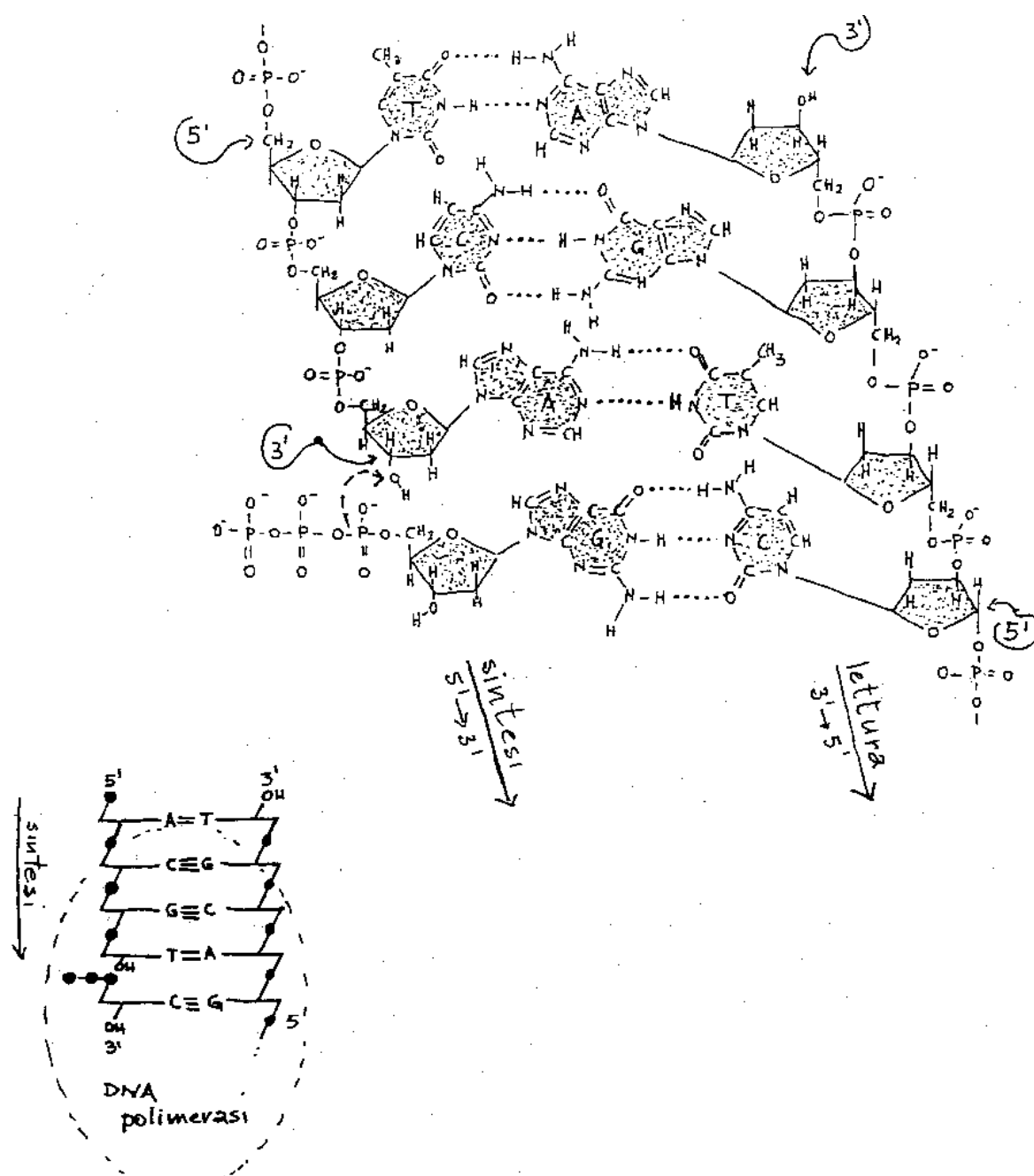



Figura 1-2. Filamento di DNAds e reazione di sintesi del DNA.
La figura mostra: un filamento di DNAds, il suo orientamento 5'-->3' definito dalla posizione dell'acido fosforico legato al carbonio in 5' (C5') del desossiribosio, la direzione di sintesi (5'-->3') della DNA-polimerasi, la direzione di lettura (3'-->5') del filamento complementare antiparallelo, e la formazione del legame fosfodiestere.
Il desossinucleotidetrifosfato avendo il fosfato legato in C5' forma il legame in posizione C3' del nucleotide ad esso precedente liberando acido pirofosforico. L'acido pirofosforico viene poi scisso dall'enzima pirofosfatasi. (ridisegnato e modificato da Watson, Hopkins, Roberts, Steitz, Weiner, Molecular Biology of The Gene, 2nd Ed. The Benjamin /Cummings).

2. Descrizione simbolica del DNAds.

```
5' GGAACCTTGATC 3' senso (codificante)
3' CCTTGGAACTAG 5' antisenso
```

La descrizione del DNA mediante la sequenza dei simboli delle basi dei nucleotidi è una semplificazione necessaria per poter leggere e scrivere la sequenza del DNA che altrimenti richiederebbero molto più spazio e tempo se il DNA fosse descritto con la formula di struttura (figura 1-2). Egualmente l'elaborazione e la conservazione negli archivi delle banche dati delle sequenze dati del DNA sarebbe più difficoltosa.
In natura i due filamenti complementari del DNAds (DNAds = DNA double strand) sono disposti con orientamento inverso (antiparallelo): l'estremità 5' di un filamento è in corrispondenza dell'estremità 3' dell'altro (figura 1-2) e così sono scritti anche nella forma semplificata. Quando si dice che un filamento di DNAss è complementare ad un altro vuol dire che i due filamenti hanno le basi complementari ed anche che sono disposti antiparallelamente.
Per convenzione il filamento di DNA scritto sopra, ha il 5' a sinistra e di conseguenza il suo complementare ha a sinistra l'estremo 3'.
Se il DNAds è di un gene o di un cDNA, il filamento in alto è il filamento senso (dall'inglese "sense" = che ha significato) o filamento codificante (coding strand).
Il filamento senso è il filamento che negli esoni di un gene e nel cDNA ha la direzione di lettura e la sequenza (con le T al posto delle U) identiche a quelle del mRNA che verrà trascritto dal quel gene e che codificherà la proteina. Il filamento 3'-5' è detto antisenso o anticodificante ed è il filamento che serve da stampo per la sintesi del mRNA. Le sequenze degli introni e delle altre parti di un gene sono scritte coerentemente con quelle degli esoni, pertanto il filamento 5'-3' di un gene includerà il filamento senso di tutti gli esoni ed in contiguità con essi i filamenti 5'-3' di tutti gli introni, del promotore e della regione non trascritta al 3'.
Queste convenzioni valgono anche se è scritta solo la sequenza delle basi del filamento 5'-3', senza avere ai suoi estremi i simboli 5' e 3':

```
GGAACCTTGATC   <--  la scrittura più semplificata della
                    sequenza nucleotidica del DNAds.
```

Dato che la complementarità di ogni singola base e l'antiparallelismo dei filamenti del DNAds sono regole sempre valide, è sufficiente scrivere solo le basi del filamento 5'-3' per indicare che sequenza è scritta nella direzione 5'-->3' e, se essa è una sequenza codificante proteine, il filamento indicato è quello senso. Complementarità ed antiparallelismo dell'altro filamento sono deducibili dalla sequenza di quello scritto.

Per semplicità di scrittura, di lettura e di conservazione negli archivi delle banche dati elettroniche viene scritta solo la sequenza 5'→3' del filamento senso dei geni e continua nelle sequenze ad esso contigue al 5' ed al 3' così come sono disposte nel cromosoma. Nella scritture provenienti dalle banche dati la sequenza nucleotidica del filamento senso è sempre contrassegnata con il segno più (+) al fine di distinguerlo da quello antisenso che è contrassegnato con il segno meno (-) quando si presenta la necessità di conoscerlo.
Anche le sequenze promotrici sono scritte come un singolo filamento 5'->3' (es. CCAATT box) anche se le proteine che interagiscono con esse (es. fattori di trascrizione) si associano al DNAds nella scanalatura maggiore prendendo contatto con i bordi delle basi accoppiate, cioè con ambedue le basi accoppiate e non con un singolo filamento (appendice b). La specificità dell'attività molecolare del DNA genomico è data dalla della forma e carica che la sequenza di basi accoppiate ha nelle scanalature della doppia elica.

3. La dimensione molecolare del DNA.
La dimensione molecolare del DNA è data come numero di b (base/basi), e non con il valore del peso molecolare generalmente usato per le altre macromolecole biologiche, perché le comuni tecniche di separazione (es. elettroforesi) separano frammenti di DNA in base alla loro lunghezza (numero di basi) e non in base al PM. Le elettroforesi su gel di poliacrilamide possono separare frammenti di DNA diversi in lunghezza per una singola base, ma non sono sufficientemente potenti da separare frammenti DNA con lo stesso numero di basi anche se i frammenti hanno una composizione diversa di basi e quindi anche un diverso PM.

4. Le sequenze consenso.
La sequenza di DNA (es. TATA box) a cui si lega specificamente una proteina (es. fattore di trascrizione) non è esattamente la stessa nel DNA di tutti gli individui di una stessa specie, uomo incluso. Anche le sequenze proteiche dotate di una particolare funzione possono essere non esattamente identiche (es. sequenze segnale di migrazione nel nucleo). In ambedue i tipi di sequenza nucleotidica ed aminoacidica, le varianti aventi la stessa funzione sono riunite in una sequenza consenso.
La sequenza consenso è la sequenza canonica (conforme ad una norma) contro la quale sono confrontate le sequenze ad essa simili. Il canone della sequenza consenso è la descrizione della base più frequentemente presente in ogni posizione delle sequenze di DNA (o di RNA) delle varianti considerate. Analogamente una sequenza consenso proteica è la sequenza che descrive i residui aminoacidici più frequentemente presenti in ogni posizione delle sequenze dei peptidi varianti considerati.
In figura è mostrato come è costruita una sequenza consenso di DNA (TATA box).

```
                             T A T A T A
                             C A T A A A
                             T A T A C A
                             T A C T A A
                             T G T A A A
                             T T T A A A
                             T A T A C A
                             T T A A C A
                             T A T A A A
                             T A T A A A
                             -----------
sequenza consenso ---->      T A T A A T
% di frequenza -------->     90 70 80 90 60 100
```

Le sequenze simili sono allineate e la sequenza consenso è costruita ponendo in ogni posizione la base più frequente nelle varie sequenze considerate.
La frequenza di una base (o aminoacido) può variare in relazione al numero di sequenze considerate per stabilire la sequenza consenso e si può verificare il caso in cui la sequenza consenso non sia rappresentata in nessuna delle sequenze allineate (in particolare quando è relativamente lunga, più di dieci monomeri) oppure che la sequenza sia identica alla maggior parte delle sequenze da cui è stata dedotta.
Il mantenimento nell'evoluzione di sequenze simili, confluenti in una sequenza consenso, suggerisce che esse abbiano una funzione a cui si attribuisce il mantenimento più o meno conservato della sequenza nell'evoluzione. Differenze di sequenza in genere risultano in differenze di ampiezza dell'attività molecolare e quindi dell'azione biologica associata alla sequenza. Ad esempio, differenze di sequenza nucleotidica di un promotore possono risultare in differenze di affinità verso uno stesso fattore di trascrizione e quindi in differenze di attivazione delle RNA-polimerasi e quindi dell'espressione genica.
La sequenza consenso è utilizzata per ricercare nella sequenza di un frammento di DNA la presenza di una sequenza con una particolare funzione. Ad esempio se si è determinata la sequenza di un nuovo gene e si vuole verificare se esso sia regolato da un particolare fattore di trascrizione, si cerca nella regione dei promotori del nuovo gene, una sequenza simile alla sequenza consenso, possibile bersaglio del fattore di trascrizione di interesse. Per un ricercatore che operi senza computer, questo confronto risulta più semplice rispetto al confronto della sequenza di interesse con ogni sequenza nota dei promotori, ai quali si lega lo stesso fattore di trascrizione; soprattutto quando la sequenza di interesse è lunga e altamente variabile nella stessa specie. Con la costruzione delle banche dati e dei programmi di gestione la ricerca delle sequenze consenso è effettuata elettronicamente ed il confronto della sequenza di interesse è fatto con tutte le sequenze note depositate nella banca dati utilizzata (vedere dopo in questo capitolo 1).
Quando una sequenza nucleotidica o proteica è conservata identica in tutti gli individui di una stessa specie, si assume che il cambiamento di un singolo

monomero (nucleotide o aminoacido) ne annulli l'attività molecolare. Le sequenze conservate identiche in una specie non sono definite consenso.

5. Direzione di sintesi degli enzimi DNA ed RNA polimerasi.
RNA-polimerasi e DNA-polimerasi sintetizzano nella direzione 5'-3' copiando il filamento 3'-5'. Dovendo ricopiare *in vitro* un frammento di DNAds ed essendo i due filamenti di DNA disposti in modo antiparallelo, la DNA-polimerasi ricopia i due filamenti scorrendo in direzioni opposte. Inoltre la DNA-polimerasi per iniziare a ricopiare il DNA necessita di un primer associato al filamento di DNA da ricopiare. In natura, i primer della DNA-polimerasi sono mRNA sintezzati dalle RNA-polimerasi che non necessitano di primer perché sono attivate da proteine nella corretta posizione del nucleotide di inizio della trascrizione (es. fattori di trascrizione). Per la sintesi i*n vitro* del DNA sono sintetizzati degli oligonucleotidi di DNAss da usare come primer per l'enzima DNA-polimerasi, pertanto per progettare e poi sintetizzare i primer occorre rispettare la regola di Chargaff e l'antiparallelismo dei due filamenti, al fine che l'enzima DNA-polimerasi possa ricopiare il filamento.

```
5' GAAGTGATTA--> sì        no <--TTGGCCTTG 3'
3' CTTCACTAATCTTTAAGTCATCGGGTTATCAACCGGAAC 5'
```

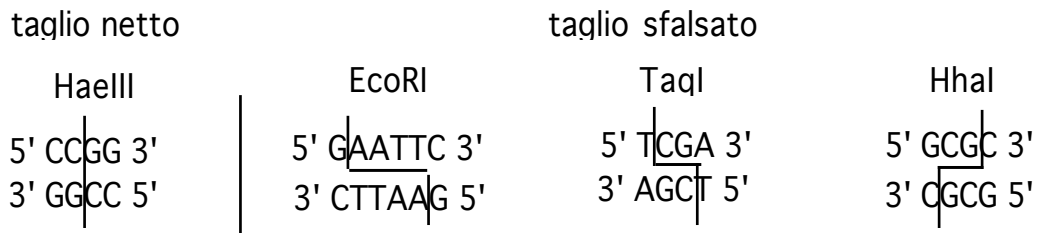
6. Siti ed endonucleasi di restrizione di tipo II.
Il DNAds può essere tagliato a livello di particolari sequenze del DNA, dette siti di restrizione, che corrispondono a sequenze brevi (4-8 coppie di basi) riconosciute da enzimi endonucleasi di tipo II (anche essi detti di restrizione) che specificamente idrolizzano i legami fosfodiestere tra due nucleotidi interni alla sequenza di restrizione. Ogni enzima riconosce un'unica sequenza di restrizione che se diversa anche per una singola base non è più riconosciuta dall'enzima (vedere capitolo 3 polimorfismo RFLP), pertanto non esistono sequenze consenso di sequenze di restrizione. Enzima e sito sono indicati con la stessa sigla che indica la specie batterica da cui è stato purificato l'enzima seguito da un numero romano che indica l'ordine cronologico con il quale più enzimi, che riconoscono sequenze diverse, sono stati purificati dallo stesso batterio. EcoRI è il primo enzima purificato da E. coli del ceppo RY13.
Le sequenze che costituiscono i siti di restrizione sono palindromi, cioè identiche nei sensi di lettura 5'--->3' dei due filamenti complementari, pertanto permettono alla endonucleasi di riconoscere i due filamenti del DNAds e quindi di tagliarli in posizioni simmetriche generando estremità nette o sfalsate.
Sono note più di 150 differenti sequenze/enzimi di restrizione. Con il taglio sfalsato uno dei due filamenti del DNAds ha un estremo più lungo, per alcuni il filamento 5'-3', per altri il filamento complementare.

taglio netto | taglio sfalsato

HaeIII
5' CCGG 3'
3' GGCC 5'

EcoRI
5' GAATTC 3'
3' CTTAAG 5'

TaqI
5' TCGA 3'
3' AGCT 5'

HhaI
5' GCGC 3'
3' CGCG 5'

Questo tipo di estremità asimmetriche sono dette coesive, perché tendono a riunire frammenti di DNA ottenuti con taglio operato da uno stesso tipo di enzima di restrizione.
Il termine "restrizione" originò dalla constatazione che alcuni ceppi batterici erano resistenti all'infezione di alcuni batteriofagi in virtù di enzimi DNA-endonucleasi che digerivano il DNA di quei batteriofagi. Mentre gli stessi batteri "restringevano" (limitavano) la possibilità ad essere infettati a quei batteriofagi che avevano il proprio DNA resistente alle loro endonucleasi che furono chiamate "endonucleasi di restrizione". I batteri modificano covalentemente il proprio DNA mediante enzimi DNA-metilasi sequenza-specifici in modo che esso non possa essere digerito dai loro stessi enzimi di restrizione. I batteriofagi, cresciuti in un dato ceppo di batteri, hanno il proprio DNA modificato come il DNA dei batteri ospitanti e quindi resistenti alle loro endonucleasi pertanto sono batteriofagi capaci (ristretti) di infettare i batteri di quel dato ceppo e non altri ceppi.
Molte specie batteriche posseggono enzimi di restrizione che sono distinti in 3 tipi. Le endonucleasi di tipo I e III sono enzimi di grosse dimensioni con più subunità ed hanno attività endonucleasica e metilasica sul DNA. Le endonucleasi di tipo I una volta che hanno riconosciuto la sequenza bersaglio si spostano a caso di 1000-5000b sul filamento di DNA (utilizzando l'energia di idrolisi dell'ATP) e dove si fermano operano il taglio del DNAds. Le endonucleasi di tipo III si spostano (utilizzando ATP) e tagliano il DNAds a circa 25b dalla sequenza di riconoscimento.
Le endonucleasi di tipo II sono il tipo di endonucleasi di restrizione più semplice, non richiedono ATP, tagliano il DNAds all'interno della sequenza di riconoscimento e data l'alta specificità del punto di taglio del DNAds che è "sequenza e base specifico", sono utilizzate nelle tecnologie del DNA ricombinante. Quando si parla genericamente di enzimi di restrizione senza dire il tipo ci si riferisce alle endonucleasi di tipo II.
Altri tipi di enzimi nucleasi (non di restrizione) sono indicati in figura 1-3.
7. Mappe di restrizione.
La formazione dei siti di restrizione nel genoma è casuale. Le mutazioni spontanee e la necessità di avere nel genoma certe sequenze nucleotidiche per svolgere certe funzioni (es. sequenze codificanti) portano casualmente alla

formazione di sequenze di restrizione. E si assume che la sensibilità alle endonucleasi di restrizione non porti alcun vantaggio selettivo per gli eucarioti. La casualità della formazione delle sequenze di restrizione è suggerita dall'osservazione che, nel DNA genomico, il numero di copie di un dato sito di restrizione è in relazione al numero di basi che la costituiscono e non in relazione alla sequenza che costituisce il tipo. La composizione in basi del genoma è circa 30% di G+C, di cui il 20% è GpC (equivale a scrivere 5'GC3') e 70% di A+T, da questi numeri si è calcolato che nel genoma umano con una sequenza di quattro basi (es. TCGA del TaqI) esista mediamente un sito ogni circa 1400b (con un totale di $2x10^6$ siti), con sei basi (GAATTC del EcoRI) un sito ogni 3100b (totale di $1x10^6$ siti). Tuttavia, la distribuzione sul DNA di una dato sito di restrizione è irregolare (le distanze in basi tra un sito e l'altro possono essere molto diverse). La posizione dei siti di restrizione su una tratto di DNA (es. gene o DNA spaziatore) è specifica di quel tratto di DNA genomico ed è detta per quel tratto di DNA mappa di restrizione di quel dato enzima, (ovviamente la mappa è diversa se si usa un differente enzima di restrizione). La mappa diviene sempre più specifica per individuare quel tratto di DNA, utilizzando più tipi endonucleasi e quindi identificando un numero maggiore di siti di restrizione. Mappe di restrizione di una stesso tratto di DNA cromosomico (stesso locus) di individui diversi di una stessa specie differiscono l'una rispetto all'altra per la presenza od assenza di uno o pochi siti di restrizione. Conoscendo la sequenza di un frammento di DNA è possibile mediante un programma di gestione ed un computer stabilire elettronicamente la mappa di restrizione di tutti i siti di restrizione presenti in quella sequenza. Questa mappa generale è utile per conoscere quali enzimi di restrizione possono essere utilizzati nella manipolazione di quel frammento di DNA (vedere capitolo 2).

Mappa di restrizione

In figura sono indicati solo i siti di taglio.

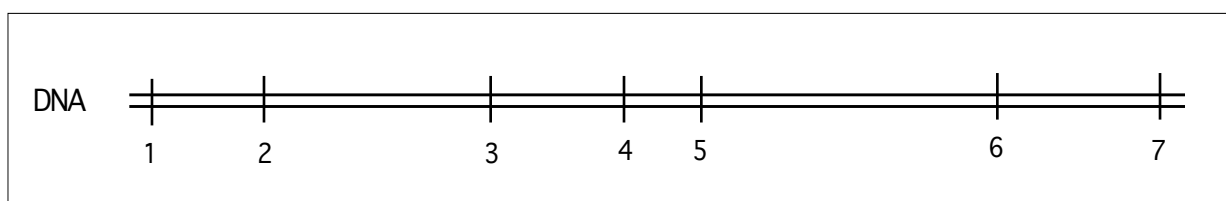
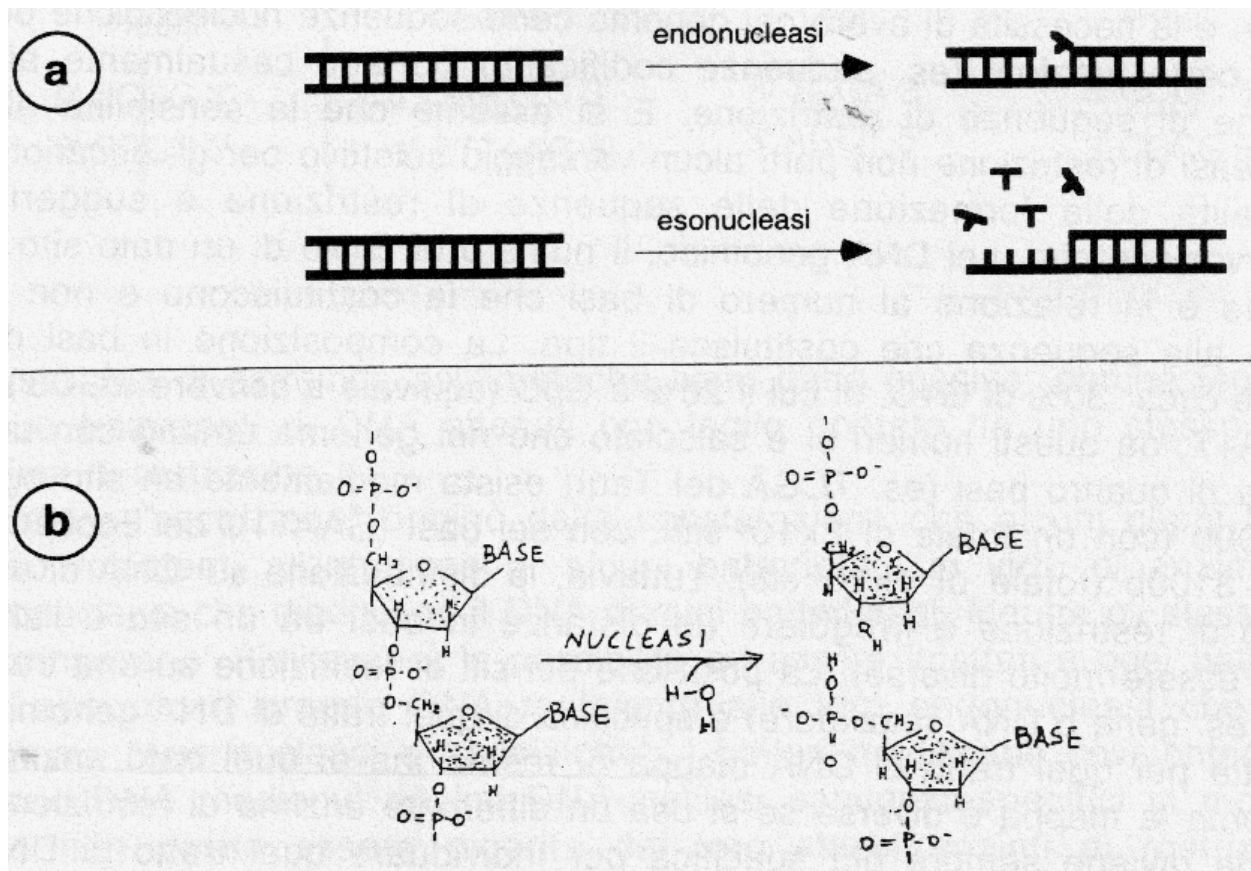



Figura 1-3. Enzimi endonucleasi ed esonucleasi.
a) reazioni schematiche delle endonucleasi e esonucleasi; b) un esempio di reazione idrolitica catalizzata da un tipo di DNAasi.
Le nucleasi sono enzimi che idrolizzano il legame fosfodiestere 5'-3' che lega i ribosi degli acidi nucleici. Questi enzimi possono essere specifici per gli RNA (Ribonucleasi = RNAasi), per il DNA (Desossiribonucleasi = DNAasi) o essere specifici solo per il legame fosfodiestere degli acidi nucleici ma non per il tipo di zucchero (ribosio o desossiribosio) per cui idrolizzano sia l'RNA che il DNA. Un'altra classificazione delle nucleasi è basata sul punto di attacco dell'acido nucleico: endonucleasi che attaccano il legame fosfodiestere all'interno della catena dell'acido nucleico ed esonucleasi che attaccano il legame fosfodiestere specificamente agli estremi 5' o 3' dell'acido nucleico liberando in sequenza nucleotidi monofosfati. I mononucleotidi liberati sono rispettivamente 5' (5'NMP) o 3' mononucleotidi (3'NMP). Le nucleasi possono scindere il legame fosfodiestere esistente tra nucleotidi di qualsiasi base, altre preferiscono i residui purinici, altre i pirimidinici ed altre ancora sono specifiche per sequenze. Le DNAasi esistono con tutti i tipi di specificità sopra elencati ed inoltre possono essere specifiche per singolo o per doppio filamento di DNA. Nell'uomo le nucleasi sono presenti nei lisosomi delle cellule di tutti i tessuti. Il pancreas secerne nell'intestino nucleasi per la digestione degli acidi nucleici contenuti negli alimenti. Le RNAasi possono essere specifiche per singoli o per doppi filamenti di RNA e la RNAasi H è specifica per filamenti ibridi DNA-RNA. Molte DNAasi ed alcune RNAasi sono utilizzate nelle tecnologie del DNA ricombinante.

8. Enzima DNA-ligasi.

Frammenti di DNA con estremità coesive complementari (prodotte dallo stesso enzima) vengono unite covalentemente con un legame fosfodiestere la cui formazione è catalizzata dall'enzima DNA-ligasi (figura 1-4). Si possono unire due frammenti o chiudere ad anello un frammento avente le due estremità coesive complementari ed identiche come disposizione (es. estremi 5' e 3' identici). Con la ligasi si possono legare covalentemente anche frammenti con estremità tagliate in modo netto.

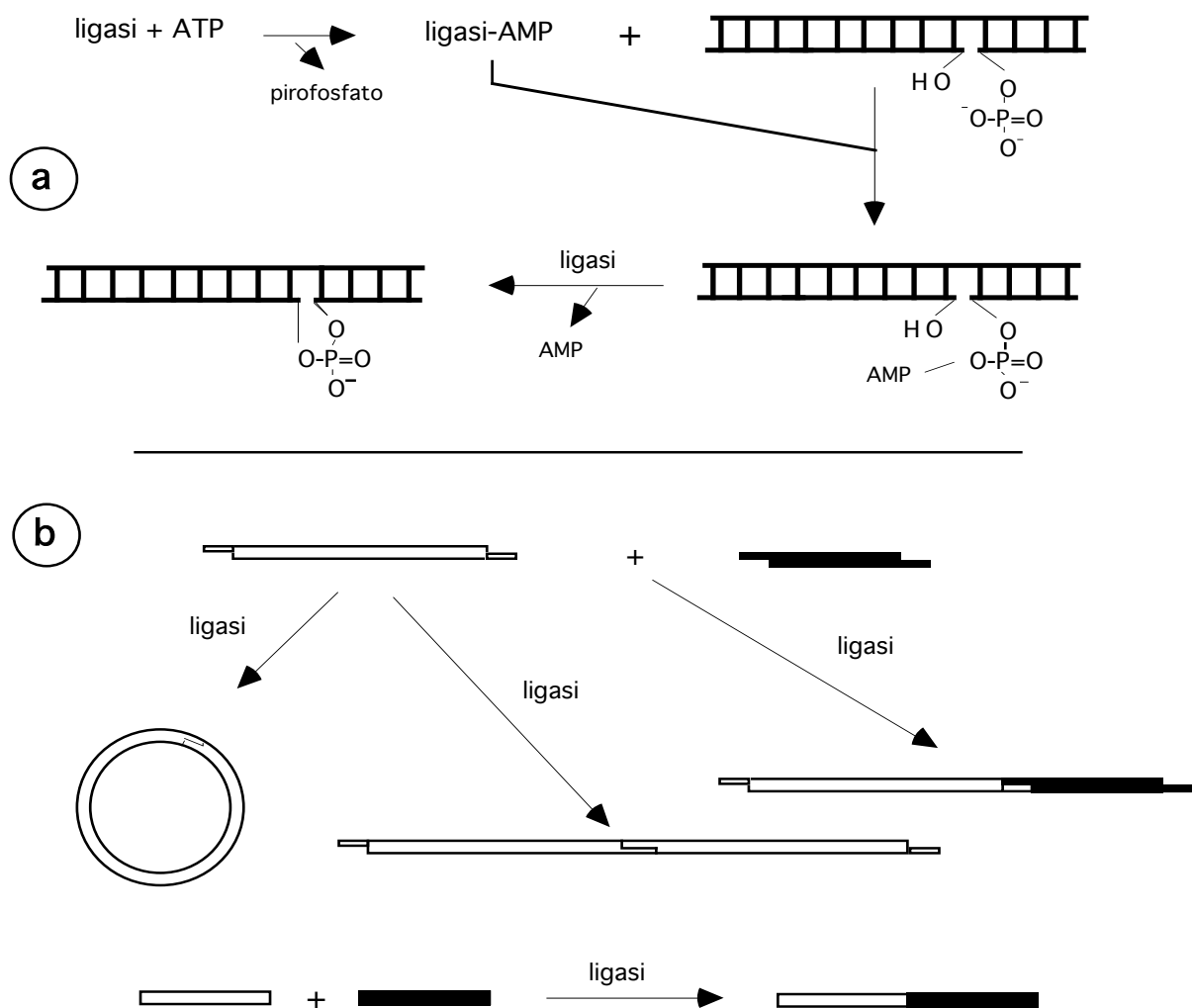



Figura 1-4. Enzima DNA-ligasi. a) Reazione della DNA-ligasi degli eucarioti; b) Substrati e prodotti della ligasi. (ridisegnato e modificato da J. D. Rawn (1989) Biochemistry, MccGraw-Hill); b) tipi di legature e ricombinazioni che possono essere fatte con la ligasi.

9. Ibridazione degli acidi nucleici.

L'associazione spontanea di filamenti singoli di acidi nucleici di diversa origine biologica o preparazione, mediante l'appaiamento delle basi complementari, è detta ibridazione. Questo è il mezzo potente ed altamente specifico per ricercare la presenza di un frammento di DNA avente una data sequenza (es. un gene o un suo frammento) tra alcuni milioni di sequenze diverse. L'ibrido può essere DNA-DNA, DNA-RNA, RNA-RNA.

Il DNAds è separato dall'agitazione molecolare, nei due filamenti se la soluzione acquosa che lo contiene viene posta a circa 95°C (non a 100°C perché l'acqua bolle e si rischia di perdere il DNA) per 1-5 minuti. La separazione del DNAds è detta denaturazione ed avviene in pochi minuti a 95°C anche se il DNAds è lungo milioni di basi.

Se la temperatura viene abbassata, i due filamenti si riassociano specificamente, riformando correttamente il doppio filamento. Questa corretta riassociazione avviene anche in soluzioni contenenti alcuni milioni di frammenti di DNA aventi sequenze diverse e presenti ciascuno in poche copie. Con l'abbassamento della temperatura ogni singolo filamento si riassocia stabilmente con il proprio complementare. I vari filamenti si muovono in soluzione venendo casualmente in contatto tra loro mossi dall'agitazione molecolare prodotta dal calore e solo quando la complementarità di sequenza è esatta, la stabilità di quel doppio filamento sarà massima ed essa renderà più difficile una nuova dissociazione. Il successivo abbassamento della temperatura stabilizza definitivamente il doppio filamento. Se alla soluzione contenente alcuni milioni di frammenti diversi di DNA aggiungiamo un frammento di DNAds (denaturato), che vogliamo usare per sondare la presenza di frammenti ad esso identici in sequenza, i suoi filamenti si ibrideranno ai filamenti ad essi complementari. Il frammento di DNAds deve essere in concentrazione sufficiente (meglio se in eccesso) per poter legare (per azione di massa) la maggior parte del DNA a lui complementare, specialmente se quest'ultimo è presente in pochissime copie. Il frammento usato per avere le caratteristiche di sonda deve essere individuabile dopo che è avvenuta l'ibridazione per poter individuare gli ibridi che si sono formati, pertanto la sonda viene marcata con isotopi o con molecole fluorescenti legate covalentemente ad essa. In passato si usavano marcare le sonde con isotopi pesanti, in questo caso, l'ibrido ha un peso inferiore a quello della sonda e superiore a quello dei frammenti naturali (non ibridi). Per cui gli ibridi se presenti erano separati e quindi individuati mediante centrifugazione in una soluzione di opportuna densità. Le sonde marcate con isotopi radioattivi sono sintetizzate sostituendo un loro atomo (es. P) con un isotopo radioattivo (es. P^{32}). L'atomo P^{32} ha tutte le proprietà chimiche dell'atomo normale, per cui non interferisce nel normale svolgimento dell'ibridazione ed inoltre emette radiazioni beta che permettono di individuare gli ibridi. Si dice che la sonda è marcata con l'isotopo P^{32}, pertanto con tale isotopo saranno marcati (quindi individuabili) gli ibridi di DNAds risultanti

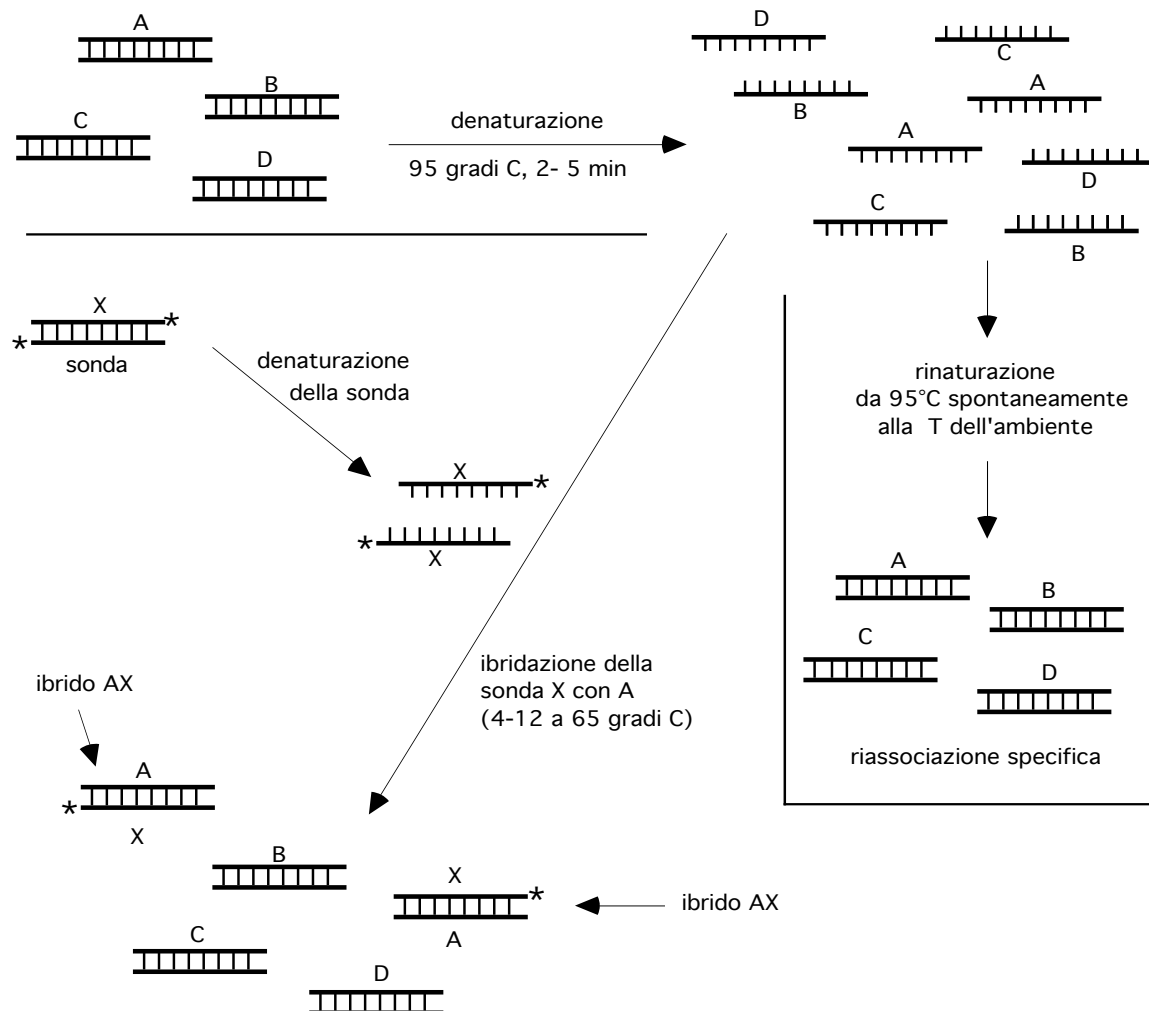



Figura 1-5. Ibridazione in fase liquida. Il DNAds è denaturato (separazione dei due filamenti) rapidamente dal calore a T vicine a 100°C. Riportando la soluzione alla temperatura ambiente la complementarità delle sequenze nucleotidiche permette la riassociazione spontanea e con altissima specificità dei filamenti di DNAss (rinaturazione), anche in miscele di alcuni milioni di filamenti diversi. L'aggiunta di un frammento di DNAds denaturato (x in figura) alla miscela di frammenti diversi di DNAds denaturati, e la successiva rinaturazione permettono di verificare se in una miscela di frammenti di DNAds aventi sequenze nucleotidiche diverse ne esiste uno che abbia una sequenza uguale (A in figura) a quella della sonda. Per poter fare questa verifica occorre che il frammento di DNA da confrontare con la miscela di frammenti di DNA sia distinguibile dagli altri cioè ad esempio marcato con isotopi (in figura è indicato con un *). Proprio perché distinguibile (marcato) il frammento viene chiamato sonda. Per semplicità si è indicata una sola copia per ogni specie molecolare di frammento di DNA (A, B, C, D, ed X) in genere la reazione di ibridazione viene effettuata in miscele di frammenti presenti in più copie. In particolare la sonda viene usata in largo eccesso quando il frammento di interesse è presente in poche copie. Per altri dettagli vedere il testo.

dall'associazione dei filamenti di DNAss della sonda complementari ai filamenti di DNAds del frammento di DNA di interesse.
Se si utilizzano sonde radioattive o fluorescenti è necessario che i frammenti di DNA da analizzare siano ancorati ad un supporto solido al fine di eliminare l'eccesso di sonde marcate che non si sono ibridate (vedere dopo analisi Southern).
Georg Hevesy è stato il primo ad esplorare l'utilità dei traccianti radioattivi in biologia, studiando nel 1923 l'assorbimento ed il movimento del piombo radioattivo nei tessuti delle piante. Rudolf Schoenheimer per primo utilizzò gli isotopi posti in posizioni specifiche delle molecole organiche per seguire le loro trasformazioni metaboliche nelle cellule.
<u>Specificità della reazione di associazione.</u>
Per avere una reazione di ibridazione con specificità assoluta (la sonda si ibrida solo alla sequenza esattamente complementare ad essa) occorre avere sonde di piccole dimensioni e condizioni sperimentali (dette di alta stringenza) in cui la temperatura di ibridazione e di lavaggio dell'eccesso di sonda (sonda che non si è ibridata), la concentrazione e composizione delle soluzioni sono scelte in relazione alla lunghezza ed alle basi della sequenza del DNA da ibridare.
<u>Temperatura di ibridazione.</u>
La temperatura di ibridazione è scelta in base alla temperatura di fusione (Tm = T of melting) dell'ibrido (sonda-DNA di interesse) alla temperatura in cui il complesso <sonda-DNA di interesse> è associato al 50%. La Tm è calcolata in relazione al numero di coppie di basi A/T e G/C. All'interazione G/C è attribuita una forza di legame doppia di quella dell'interazione A/T. Se una sonda contiene molte coppie G/C la sua Tm sarà più alta di quella di una sonda in cui prevalgono coppie A/T, cioè occorrerà una temperatura più alta (agitazione molecolare più intensa) per mantenere dissociato il 50% dei filamenti della sonda e quelli del DNA di interesse. Ibridare alla Tm è un compromesso tra avere un sufficiente numero di ibridi per operare l'indagine analitica e l'assoluta specificità. Se si operasse a $T<Tm$ si formerebbero più complessi ma si potrebbero formare anche ibridi con una o più basi disaccopiate tuttavia sufficientemente stabili per l'agitazione molecolare prodotta da una temperatura più bassa della Tm. Operando a $T>Tm$ si riduce il numero di ibridi e ciò riduce la sensibilità dell'analisi (occorre usare molte più copie di sonda e/o di DNA di interesse per avere un numero sufficiente di ibridi che possano essere rivelati. Talvolta, in relazione alla sequenza del frammento di DNA di interesse, che tende ad assumere conformazioni non favorevoli all'associazione e/o in presenza di frammenti quasi identici, occorre operare a $T>Tm$.
Le condizioni di ibridazione possono essere migliorate aggiungendo opportune quantità di denaturanti (favoriscono lo svolgimento dei filamenti) o di sali (favoriscono l'associazione del complesso, sottraggono acqua).
<u>Sonda.</u> In condizioni di buona stringenza, sonde di 15-25 basi si associano stabilmente a frammenti di DNA aventi sequenze totalmente complementari ad esse, ma non rimangono associate se c'è anche una sola base disaccoppiata.

Sonde più lunghe, dato il maggior numero di legami a H che possono formare, formano ibridi stabili anche se hanno una o più basi disaccoppiate.
Inoltre è stato calcolato che una sonda di 20b sia sufficientemente lunga da poter individuare specificamente una sequenza unica del DNA genomico umano e quindi un unico locus (posizione subcromosomica).
4^{16} (4, base della potenza, è dato dal numero delle basi del DNA) corrisponde ad oltre 4 miliardi e rappresenta il numero di sequenze diverse di 16b che allineate una dopo l'altra costituiscono una sequenza di lunghezza superiore a quella del genoma umano. Quindi teoricamente una sonda di 16 basi, qualsiasi sequenza essa abbia (sequenze ripetute escluse), ha una sequenza che è identica ad una sequenza unica del genoma umano. L'esperienza ha indicato che nel genoma umano ci sono ridondanze di brevi sequenze pertanto la lunghezza minima di una sonda è stata portata a 20 basi per avere la sicurezza che la sonda ibridi con una sequenza unica del genoma umano e quindi individui un unico locus dello stesso genoma. Questa è la teoria che è generalmente corretta, tuttavia l'unicità nel genoma delle sequenze di 20b deve e viene sempre verificata con analisi molecolari (ibridazione e sequenza dell'ibridato figura 1-6, PCR e sequenza dell'amplificato figura 1-8.). Se utilizziamo una sonda di 20b con sequenza identica ad una sequenza ripetuta nel genoma (VNTR, SINES e LINES, geni ripetuti) essa ibriderà su più loci, mentre una sonda di 20b con sequenza diversa dalle sequenze ripetute sarà unica ed ibriderà su un unico locus.
L'ibridazione per scopi analitici: individuare un frammento di DNA di interesse tra milioni di frammenti diversi (come nell'analisi di laboratorio per la ricerca delle mutazioni genetiche umane) è possibile solo se in precedenza si è determinata l'esatta sequenza del DNA di interesse con la tecnologia di Sanger (figura 1-10). Altrimenti non si può costruire una sonda con la stessa sequenza del DNA di interesse, stabilire la Tm, la T di ibridazione.

Sommario:
Per studi della genomica umana l'uso di sonde di 20basi è legato a tre caratteristiche importanti:
- a temperature di ibridazione uguali alla Tm, la sonda si lega specificamente e stabilmente ad un frammento di DNA avente una sequenza interamente complementare ad essa. Il frammento di DNA può essere anche molto più lungo della sonda purché contenga l'intera sequenza complementare alla sonda. Quando la T della soluzione è lasciata scendere alla T dell'ambiente sarà presente una quantità di ibrido sufficiente per poter valutare l'avvenuta ibridazione.
- a temperature di ibridazione uguali alla Tm, è sufficiente che sonda e DNA di interesse abbiano le rispettive sequenze diverse per una singola base affinché il loro ibrido <sonda-frammento di DNA di interesse> non sia stabile.
20b è un numero di basi sufficiente per individuare una sequenza unica del genoma umano.

Ibridazione per individuare il DNA del quale non è conosciuta l'esatta sequenza. L'uso di sonde aventi più di 20b e condizioni di ridotta stringenza sono utilizzate quando si vogliono ricercare frammenti di DNA dei quali non conosciamo ancora l'esatta sequenza, pertanto utilizziamo sonde con più di 20b assumendo la presenza di uno o più disaccoppiamenti nell'ibrido. Un esempio può essere l'uso come sonda di un frammento di DNA di un gene di topo del quale conosciamo la funzione al fine di ricercare il gene omologo nel genoma umano (vedere dopo vaglio delle genoteche). Assumendo che possano esserci variazioni di sequenza tra i due geni conseguenti l'evoluzione delle due specie, conviene utilizzare una sonda più lunga al fine di ottenere un ibrido che non si formerebbe con una sonda più breve ed in condizioni di alta stringenza. Data la minore specificità, c'è il rischio di ottenere ibridi di geni umani non omologhi del gene di topo, tuttavia analizzando la completa sequenza del gene umano individuato in questo modo si può accertare o meno l'omologia (vedere dopo). L'alta specificità dell'ibridazione, cioè l'unicità del corretto appaiamento dei due filamenti di DNA dipende dal fatto che due sequenze nucleotidiche complementari hanno le basi capaci di formare solo due (coppie A/T) o tre (coppie G/C) legami ad idrogeno. Inoltre per la legge di Chargaff, le coppie di basi si formano sempre, perché più stabili, tra una purina (G o A) ed una pirimidina (C o T). Date le diverse dimensioni delle purine (due anelli: esa- e penta-atomico) e delle pirimidine (un anello: esa-atomico), questi accoppiamenti (purina-pirimidina) determinano che il filamento di DNAds abbia una costante dimensione trasversale, che è la più stabile. Le coppie di basi e quindi i relativi legami ad idrogeno sono posti in piani perpendicolari all'asse della doppia elica, per cui la specificità dell'appaiamento di due filamenti di DNA dipende dalla specifica sequenza di coppie e triplette di legami ad idrogeno. La specificità dell'appaiamento delle singole coppie di basi (AT e GC) dipende dalla precisa disposizione dei gruppi atomici capaci di formare due o tre legami a idrogeno e dall'energia libera liberata nella formazione di questi legami, che è sempre superiore a quella liberata da accoppiamenti diversi da AT e GC. Le coppie e le triplette di legami a idrogeno sono le unità di riconoscimento molecolare per il corretto appaiamento delle basi e quindi dei filamenti di DNA.

Nota. I due filamenti di un frammento di DNAds di solo 200 basi, sono associati tra loro con una forza di circa 500kC/mole data dai legami a H (idrogeno) tra le basi (e di altre interazioni che contribuiscono all'associazione). 500kC è una forza superiore a quella dei legami covalenti (50-150kC), tuttavia l'energia dell'agitazione molecolare a 95°C in pochi minuti è capace di dissociare i due filamenti di DNA ma non a spezzare i legami covalenti che legano i loro atomi.
La minore resistenza alla dissociazione del DNAds è data dal fatto che i due suoi filamenti sono tenuti associati da circa 500 legami a idrogeno aventi una forza media di circa 1kC/mole. Si assume che l'agitazione molecolare a 95°C inizialmente dissoci 10-20 legami a H ad un estremo del DNAds e mantenga separati i filamenti, indebolendo così la stabilità del complesso (DNAds). Nel tempo, la stessa agitazione molecolare provocherà la rottura di altri legami a H fino a dissociare completamente il DNAds. I filamenti complementari tenderanno a riassociarsi, ma l'energia dell'agitazione molecolare a 95°C li manterrà separati data la debolezza dei legami a quella temperatura.Alcuni autori considerano la tendenza ad associarsi tra molecole, misurata come affinità, equivalente all'amore tra gli esseri viventi e la specificità molecolare ad una rispettata monogamia. In questo caso le basi molecolari dell'amore sarebbero chimico-fisiche.

10. Analisi Southern.

L’analisi Southern è l’ibridazione su supporto solido di frammenti di DNA frazionati mediante elettroforesi (figura 1-6). Il DNA umano digerito completamente con un enzima endonucleasi di restrizione come TaqI genera circa 2 milioni di frammenti di lunghezze diverse. I frammenti generati con la digestione del DNA sono gli stessi (stessa sequenza e lunghezza) per il DNA di uno stesso individuo; per individui diversi di una stessa specie possono esistere differenze di lunghezza dei frammenti a causa del polimorfismo che altera la sequenza dei siti di restrizione (vedere capitolo 3 ed appendice D). Se tra questi vogliamo ricercare quello che ha la sequenza che ci interessa, inizialmente è conveniente frazionare i vari frammenti sulla base della loro dimensione mediante elettroforesi su gel di agarosio (carboidrato polimerico ad alto PM). L'agarosio idratato, in relazione alla sua concentrazione (intorno al 1%), costituisce una rete molecolare tridimensionale. Attraverso le maglie (pori) di questa rete passano i frammenti di DNA. I frammenti più piccoli passano più facilmente e al termine dell'elettroforesi si troveranno più vicini al polo positivo. I frammenti con un più alto numero di basi migrano più lentamente per la maggiore difficoltà a passare tra le maglie dell'agarosio, ed essi si troveranno più vicini al polo negativo, cioè più vicini al pozzetto (origine dell'elettroforesi) dove sono stati posti sul gel, poco prima di aprire il passaggio della corrente. I filamenti di eguale dimensione tenderanno ad ammassarsi tutti in una stessa zona discreta del gel anche se hanno sequenze diverse. Se il DNA viene reso visibile appariranno nel gel delle piccole strisce dette <u>bande</u> (ordinate e perpendicolari alla direzione della corrente), ciascuna di esse contiene frammenti di DNA di una stessa dimensione (numero di basi) ma non necessariamente della stessa sequenza (pertanto una stessa banda può contenere frammenti aventi sequenze diverse). Tutti i frammenti di DNA migrano verso il polo positivo perché il DNA ha una carica fortemente negativa. Questo perché ogni singolo nucleotide del DNA ha un gruppo acido (dell'acido fosforico) libero (non esterificato al desossiribosio). Al pH (circa 8) del tampone usato per l'elettroforesi, il gruppo è dissociato (ha perso un H^+) e quindi carico negativamente. Risulta così che il rapporto tra la carica negativa e dimensione dei vari frammenti di DNA è costante ed in conseguenza di ciò, nel campo elettrico, i frammenti sono trascinati da una forza proporzionale al loro numero di basi. Se l’elettroforesi fosse eseguita in un liquido ideale (senza maglie di agarosio, né attriti), i frammenti di DNA migrerebbero tutti (grandi e piccoli) con la stessa velocità percorrendo nello stesso tempo la stessa distanza (come le sfere di ugual diametro, ma di peso diverso, che lanciate da Galileo dalla torre di Pisa arrivavano a terra insieme). La presenza dei pori dell'agarosio causa la separazione dei frammenti di DNA in relazione alle loro dimensioni. Sul gel di agarosio, in un pozzetto vicino a quello contenente il campione di DNA da analizzare, viene posta una soluzione contenente frammenti di DNA radioattivo di varia e nota lunghezza (marcatori di

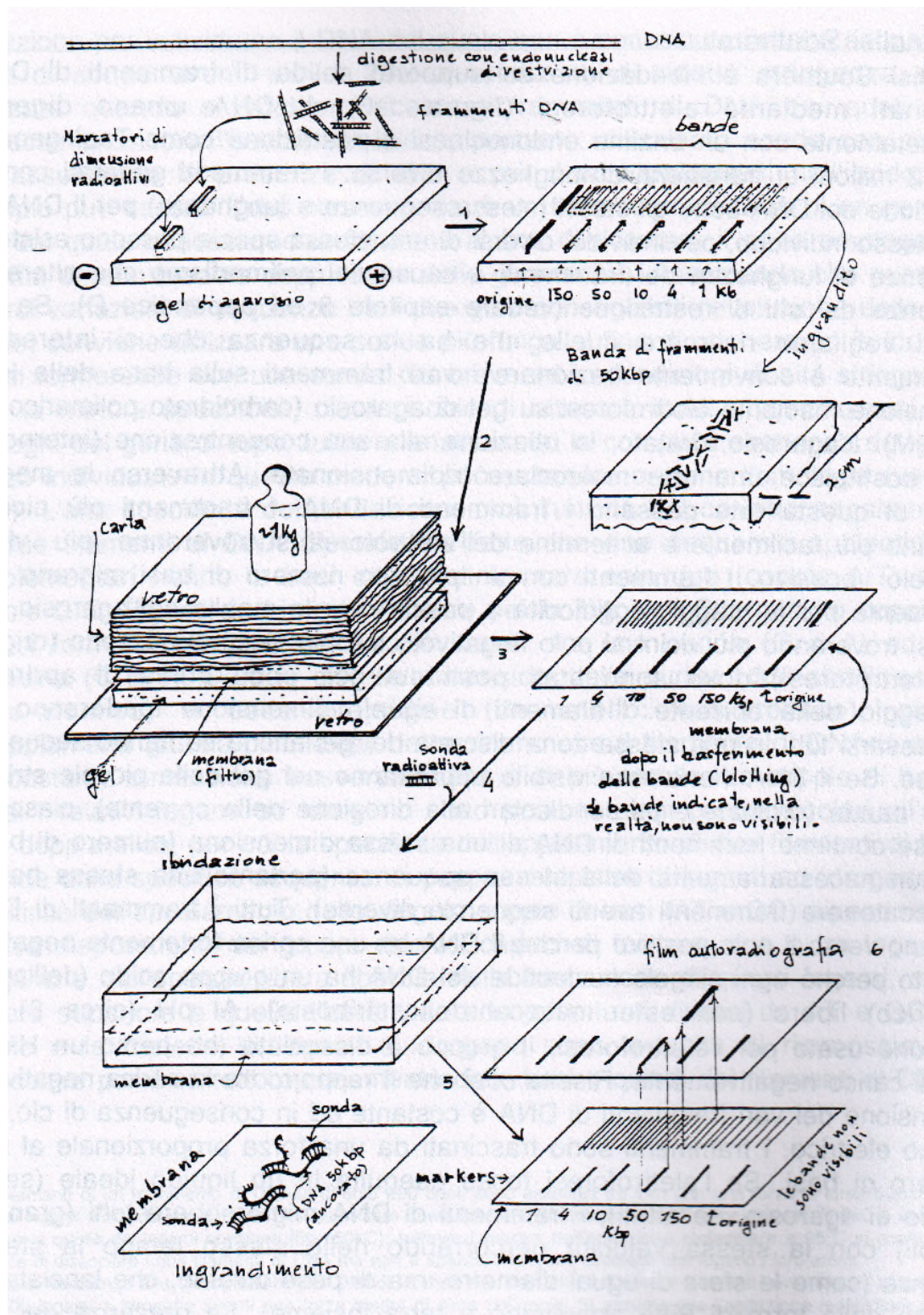



Figura 1-6. Analisi Southern. Frammentazione del DNA (digestione con enzima endonucleasi di restrizione), trasferimento del DNA dal gel di elettroforesi su un filtro di nitrocellulosa (Southern blotting), ibridazione con sonda marcata (radioattiva), autoradiografia. (ridisegnato e modificato da Recombinant DNA, Watson JD, Gilman M, Witkowski J, Zoller M. (1992) Scientific American Books, 2nd ed.).

dimensione). Dopo avere eseguito l'elettroforesi, i marcatori permettono di valutare con buona approssimazione la dimensione del frammento di DNA contenente la sequenza di interesse ed anche di poter comparare elettroforetogrammi diversi per tempi di elettroforesi (migrazioni del frammento più o meno estese), composizione del gel. Per individuare in quale banda si trovi il frammento di DNA di interesse bisogna operare una ibridazione con una specifica sonda, tuttavia ciò non può essere fatto direttamente sul gel (è appiccicoso e delicato). Si opera quindi il trasferimento del DNA dal gel su un sottile supporto solido (filtro di nitrocellulosa o membrana di nylon) in modo che la disposizione che le bande avevano nel gel sia esattamente mantenuta sul filtro. Il trasferimento è detto Southern blotting. Southern è il cognome del ricercatore che ha ideato e realizzato la tecnica di trasferimento (blotting = "macchiatura" del filtro con i frammenti di DNA). Il filtro e la membrana sono filtri molecolari, hanno dei micropori di dimensioni sufficientemente piccole da permettere il passaggio dell'acqua e soluti ma non quello di frammenti anche piccoli di DNA o RNA. Si pone il filtro (di dimensioni uguali a quelle del gel) sul gel stesso, poi sul filtro sono posti 5-6 cm di fogli sovrapposti di carta assorbente.

Sopra tutto si pone un peso di circa 1kg. Per capillarità, il liquido contenuto nel gel si muove verso la carta e, attraversando i micropori del filtro, trascina con se anche i frammenti di DNA che però sono trattenuti dalla membrana, rimanendovi adsorbiti (ancorati). La membrana viene posta ad ibridare con la sonda radioattiva per 4-12h. Al termine dell'ibridazione la membrana è lavata dell'eccesso della sonda radioattiva (molecole di sonda che non si sono ibridate e quindi non rimangono associate al DNA né al filtro), asciugata e su essa, in camera oscura, è posto un film per radiografie. Le radiazioni emesse dagli ibridi "sonda-frammento di DNA" convertono i sali di argento del film in argento metallico che è nero causando l'autoradiografia. Dalla presenza sul film di una banda nera autoradiografica, si ha l'indicazione che nel DNA analizzato c'è un frammento di DNA con sequenza identica (o in relazione alle condizioni di ibridazione, molto simile a quella della sonda). Dalla posizione della banda, relativamente a quelle dei marcatori di dimensione, si ha l'indicazione della dimensione del frammento individuato. Questa metodica è indicata come analisi Southern. Nella banda, in cui si trova il frammento di DNA con la sequenza di interesse, possono esserci anche frammenti di DNA della stessa dimensione ma che hanno sequenze nucleotidiche diverse. Per purificare il frammento di interesse dagli altri frammenti di DNA, occorre estrarlo dal gel e poi subclonarlo (figure 1-7 e 1-8).

La tecnica Southern è molto potente, può rivelare una sequenza presente in singola copia nel genoma analizzando frammenti di DNA estratto da 1mm^3 di cellule (circa 10^6 cellule). Circa $2x10^6$ copie di uno stesso frammento da individuare mescolato ad alcuni milioni di frammenti diversi anch'essi presenti in milioni di copie. I numeri di molecole appaiono grandi (circa 2 milioni) sono solo circa 3 millesimi di femtomole (3×10^{-18} moli) una quantità in peso estremamente piccola. Se il frammento di DNAds è lungo 1.500b ha un PM di

circa 900kg, due milioni delle sue molecole pesano circa 0,3 femtogrammi ($3x10^{-15}$ grammi). Classica ricerca di un ago in un pagliaio di aghi molecolari !

11. Ibridazione del DNA su spot (spot=macchia/posto).
Per verificare rapidamente la presenza di un frammento di DNA tra milioni di altri frammenti (ad esempio ottenuti mediante digestione del genoma umano con uno o due enzimi di restrizione) è sufficiente depositare un poco della soluzione contenente il DNA genomico digerito su un filtro di nitrocellulosa e quindi lasciare evaporare l'acqua. Con una sonda si esegue l'ibridazione direttamente sullo spot, si lava via l'eccesso di sonda che non si è legata al DNA e poi con l'autoradiografia o un contatore di radioattività si misura la radioattività contenuta nello spot. La presenza di radioattività indica che la sonda si è ibridata al DNA contenuto nello spot.
L'analisi Southern è più completa di quella su "spot" perché, in conseguenza del frazionamento operato con l'elettroforesi, indica oltre la presenza, anche la dimensione del frammento e questa è una caratteristica importante, perché permette di confrontare la costanza dell'integrità e della quantità del frammento durante processi di purificazione, e la sua frammentazione dopo l'esposizione ad enzimi di restrizione.
Un'altra importante applicazione dell'analisi Southern è la tecnica per la determinazione dell'impronta del DNA (figura 3-15).

12. Analisi Northern.
La stessa metodologia Southern è utilizzata per analizzare miscele di mRNA e così anche di altre specie di RNA. Gli mRNA da analizzare sono usati come tali perché hanno dimensioni sufficientemente piccole da poter essere frazionati mediante elettroforesi senza dover procedere alla digestione con enzimi di restrizione che è necessaria per il DNA genomico. L'analisi degli mRNA è stata chiamata Northern per distinguerla da quella del DNA (analisi Southern).

13. Analisi Western.
Una tecnica simile all'analisi Southern è utilizzata per individuare una proteina tra miscele di proteine. Le proteine sono frazionate mediante elettroforesi su gel contenente Na-dodicilsolfato (SDS). La parte idrofobica (dodicil) del SDS si associa con legami idrofobici alla proteine (circa il 50% dei residui aminoacidici di una proteina sono idrofobici) e la parte carica negativamente (solfato è carico negativamente perché il legame salino con il Na si dissolve nell'acqua del gel) permette la migrazione delle proteine nel campo elettrico. Le maglie del gel di poliacrilamide frazionano le proteine in base alla loro dimensione molecolare. Le proteine legate SDS sono trasferite dal gel al filtro mediante elettroforesi (electroblotting) e la ricerca di interesse è fatta con anticorpi che associano specificamente alla proteina. Il nome Western è stato scelto per distinguere questa tecnica dalle altre due usando ancora un aggettivo derivato da un punto cardinale.

Tecnologie per la sintesi di DNA

Subclonazione del DNA.

La subclonazione del DNA è la clonazione di un frammento di DNAds in un vettore in genere diverso da quello in cui era stato clonato precedentemente. La clonazione del DNA è l'isolamento di un frammento di DNA da un insieme di frammenti diversi, come ad es. i frammenti originati dalla digestione con un enzima di restrizione del DNA genomico (figure 1-13 e 1-14). Isolato con la clonazione un frammento di interesse si sottopone a subclonazione al fine di verificare la sua purezza e/o produrne grande quantità. Il frammento da subclonare è un frammento puro, noto nella sequenza, ed i suoi estremi sono stati tagliati con un enzima di restrizione noto. La tecnologia concettualmente è molto semplice: il frammento di DNAds di interesse è legato covalentemente al DNAds di un vettore che in genere è un plasmide (figura 1-7). Il frammento di DNA legato al DNA del vettore è detto costrutto: [DNA da sintetizzare]-[DNA del vettore]. Il DNA del costrutto è detto ricombinante, perché mediante operazioni *in vitro* sono stati uniti frammenti di DNA che in natura erano separati. Il costrutto è mescolato con una sospensione di batteri detti competenti, perché capaci di ricevere il plasmide e di farlo replicare molte volte durante una loro singola replicazione cellulare (figura 1-7b). La temperatura della sospensione viene innalzata per indurre l'ingresso del DNA-costrutto nelle cellule. L'assunzione di DNA dall'ambiente da parte delle cellule batteriche è chiamata trasformazione ed i batteri trasformati sono organismi geneticamente modificati (OGM). La sospensione di batteri ancora calda è posta, e poi gelifica, su un terreno solido (figura 1-7b) contenente nutrienti per alimentare i batteri. I batteri contenenti il costrutto aumenteranno in numero formando colonie e aumenterà anche il numero di copie del costrutto all'interno delle stesse cellule. In questo modo un grande numero di copie del vettore-inserto sarà stato replicato ed esso potrà essere riestratto dai batteri ed utilizzato per analisi ed ulteriori manipolazioni. I plasmidi sono piccole porzioni circolari di DNAds extracromosomico capaci di accettare e replicare frammenti di DNA esogeno. I plasmidi usati per le subclonazioni e clonazioni sono stati opportunamente modificati con aggiunta del DNA di geni o di DNA avente sequenze responsabili di funzioni specifiche (figura 1-7a) al fine di permettere l'inserimento del DNA di interesse nel vettore e di operare correttamente la subclonazione (figura 1-7b).

Il DNA del vettore-plasmide modificato in genere include:

a) una origine di replicazione batterica (ori) che per ogni singola replicazione del batterio ospite fa replicare centinaia di volte il costrutto favorendo un'alta produzione dell'inserto di interesse in esso contenuto.

b) un tratto di DNA, detto polylinker, in cui sono allineate le sequenze di più tipi di siti di restrizione. Con il polylinker, lo sperimentatore ha la possibilità di legare allo stesso tipo di vettore, frammenti di DNA tagliati con enzimi di restrizione diversi. Ciò risulta molto utile perché il frammento di DNA da clonare può avere al suo interno più di un sito di restrizione per cui, ai fini della subclonazione e ricombinazione *in vitro* è necessario scegliere un enzima di restrizione che non

lo tagli al suo interno. Il vettore viene tagliato con lo stesso enzima di restrizione con cui è stato tagliato il DNA di interesse in modo che essi abbiano i loro estremi complementari e quindi coesivi (figura 1-7b). Poi il DNA del vettore ed il DNA di interesse sono legati covalentemente con una reazione catalizzata dall'enzima DNA-ligasi (figure 1-7b e1-4).
c) due geni per favorire la selezione delle cellule batteriche contenenti il costrutto. Uno di questi geni conferisce ai batteri la resistenza all'antibiotico ampicillina (gene Amp^R). Il gene Amp^R codifica per un enzima che catalizza una reazione covalente sulla molecola dell'antibiotico inattivandola. Per la presenza della sequenza ori nel plasmide, il plasmide, e quindi il gene $Amp^{R,}$ è presente in grandi quantità in un singolo batterio conferendogli un'alta resistenza all'antibiotico.
L'altro (gene LacZ) codifica una parte aminoterminale della proteina enzimatica beta-galattosidasi (LacZ) che, quando sintetizzata, si assocerà ad un'altra parte dello stesso enzima sintetizzata nei batteri competenti. Le due parti dell'enzima LacZ associate sono cataliticamente attive su un substrato sintetico: Xgal (galattosio modificato) che diviene colorato blu con la reazione catalizzata dall'enzima beta galattosidasi. Xgal è mescolato al terreno di coltura e le colonie dei batteri contenenti il plasmide senza inserto (che sintetizzano l'enzima beta-galottisadasi attiva) si colorano di blu (figura 1-7b). Mentre le colonie dei batteri, che contengono il plasmide con l'inserto, non sintetizzano l'enzima beta-galattosidasi attivo perché l'inserimento del DNA di interesse nel polylinker altera la codificazione dell'enzima. Queste colonie rimangono non colorate e quindi distinguibili dalle altre.
I batteri competenti ed i plasmidi sono prodotti dell'ingegneria genetica: ori, siti di restrizione, gene Amp^R e il frammento del gene lacZ provengono da DNA di cellule e plasmidi diversi.
I due diversi tipi di vaglio delle colonie batteriche permettono di selezionare decine di colonie contenenti il plasmide con l'inserto di interesse.
Il doppio vaglio appare necessario se si considera che il frammento di DNA di interesse viene fatto reagire in presenza di ligasi con un eccesso di plasmide-vettore (al fine di legare la maggiore quantità di DNA di interesse) per cui alla fine della reazione avremo più molecole di vettore che di costrutto (vettore-inserto ciclizzato). Questo soluzione, contenente vettori e minori quantità di costrutti, verrà fatta interagire con milioni di batteri al fine di avere un'alta probabilità che i relativamente pochi plasmidi ricombinanti (costrutti) siano inseriti nel maggior numero possibile di batteri. Pertanto il vaglio delle colonie batteriche con l'antibiotico elimina l'eccesso di batteri non transfettati (senza gene Amp^R i batteri muoiono); il vaglio delle colonie con Xgal permette di selezionare le colonie che includono solo il plasmide (colonie blu, beta-galattosidasi positive) da quelle che includono il costrutto (colonie non colorate di blu, beta-galattosidasi negative).
Da una colonia non-colorata con uno stecchino da denti sterile, viene prelevata una piccola quantità di batteri (teoricamente uno, per essere ancora più sicuri di avere una colonia omogenea, se per caso nella colonia ci fosse qualche batterio)

e fatta crescere in un terreno di coltura liquido che permette una crescita più rapida di quella che avviene nei terreni solidi. Raggiunta una ottimale quantità di batteri, essi vengono raccolti mediante centrifugazione e distrutti al fine di estrarre da essi il DNA del costrutto. Il costrutto viene digerito con l'enzima di restrizione, lo stesso che era stato utilizzato per tagliare il plasmide, ed il frammento di interesse isolato dal DNA plasmidico mediante elettroforesi.
La doppia selezione dei batteri mediante i prodotti dei geni Amp^R e lacZ permette di ottenere il frammento di interesse inizialmente presente in pochissime copie di plasmide. Come in natura la selezione favorisce la crescita dei pochi più atti a crescere nelle condizioni ambientali.

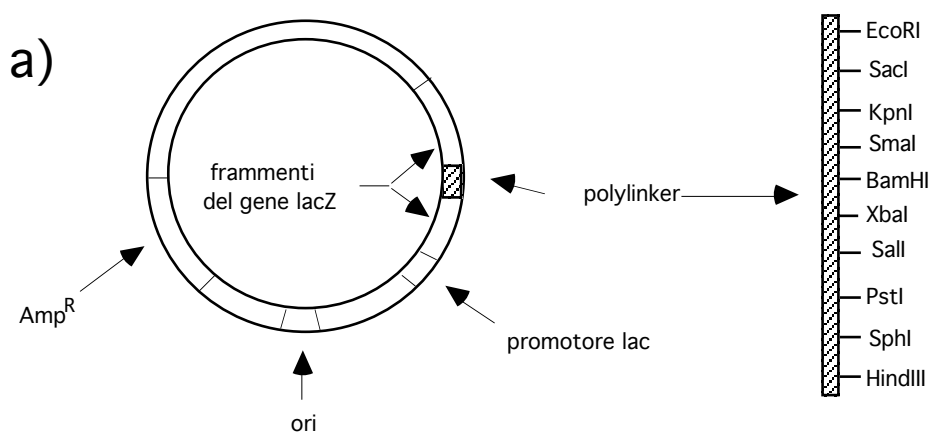



Figura 1-7. Subclonazione di un frammento di DNAds.
a) Plasmide. Il plasmide include: un'origine di replicazione (ori) che permette la replicazione del plasmide in centinaia di copie all'interno di ogni cellula; il gene Amp^R, che conferisce la resistenza all'antibiotico ampicillina e permette di selezionare positivamente i batteri che contengono il plasmide impedendo la crescita dei batteri che non lo contengono (figura 1-7b); il polylinker, sequenza contenente vari siti di restrizione. Il polylinker è inserito all'interno di un frammento del gene lacZ che codifica la parte amino terminale dell'enzima beta-galattosidasi in modo da non alterare il quadro di lettura del gene dell'enzima.
b) Pagina seguente. Subclonazione di un frammento di DNA in un plasmide.
(ridisegnato e modificato da figura 7-1 di Recombinant DNA, Watson JD, Gilman M, Witkowski J., Zoller M. Scientific American Books, 2nd ed. e figura 4.9 di Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK).

enzima di
restrizione
taglio
polylinker
Amp^R
DNA
da inserire
enzima ligasi
legatura
plasmide che è stato
richiuso senza inserto
DNA
ricombinante
b)
introduzione del plasmide nelle
cellule batteriche competenti
cellula che non ha
ricevuto il plasmide
1a selezione
Crescita delle cellule batteriche su piastre di terreno solido
contenente Ampicillina. Solo le cellule contenenti il plasmide
(con o senza inserto) cresceranno formando colonie
2a selezione
Crescita delle cellule batteriche su piastre di terreno solido contenente un composto
incolore (X-gal). L'enzima beta-galattosidasi catalizza la reazione di conversione del X-gal
in un composto insolubile blue. Pertanto le cellule contenenti il plasmide senza inserto
formano un polipeptide che porta alla beta-galattosidasi attiva e quindi alla formazione di
colonie colorate di blue. Le colonie delle cellule contenenti i plasmidi con l'inserto (cloni
positivi) sintetizzeranno un polipeptide inattivo e risulteranno bianche.
clone positivo

Tecnica della reazione a catena della DNA-polimerasi

La tecnica della reazione a catena della DNA-polimerasi (PCR = Polymerase Chain Reaction) permette di sintetizzare frammenti di DNAds di circa 100b fino a circa 40kb, utilizzando come stampo DNAds (figura 1-8). Per attuare la reazione, vengono sintetizzati due desossioligonucleotidi di DNA a singolo filamento (DNAss) di 15-20b mediante un'apposita macchina sintetizzatrice. La sequenza degli oligonucleotidi è fatta su progetto dello sperimentatore (non per ricopiatura di uno stampo di DNA) in modo che un oligonucleotide sia esattamente complementare ad una regione posta all'estremo 3' del filamento senso e l'altro complementare alla regione posta all'estremo 3' del filamento non senso del frammento di DNA di interesse. I desossioligonucleotidi sono i primer che l'enzima DNA-polimerasi utilizzerà per sintetizzare l'intero filamento di DNAds. La DNA-polimerasi (Taq-polimerasi) è purificata dal batterio *Thermus aquaticus* che vive nelle acque termali a 75°C, per cui l'enzima è molto resistente al calore (molto più degli enzimi estratti dai batteri che vivono a T ambientali più basse).
I due primer, Taq DNA-polimerasi, deossinucleotiditrifosfati (dATP, dCTP, dGTP e dTTP) e DNAds da ricopiare sono mescolati in una soluzione tamponata (la reazione di sintesi del DNA libera acido pirofosforico che se non tamponato acidificherebbe la soluzione inibendo la sintesi) e posti in una microprovetta (il volume della soluzione di sintesi può essere anche pochi μl). Una macchina (termociclizzatore) ciclicamente alza la temperatura della soluzione a circa 95°C per denaturare il DNA stampo, quindi abbassa la temperatura alla Tm dei primer (50-70°C, figura 1-8) per permettere l'associazione dei primer al DNA stampo e poi la alza nuovamente fino a 72°C, temperatura ottimale per l'enzima Taq-polimerasi, che iniziando dai due primer, catalizza la sintesi del DNA sui due filamenti del DNA di interesse (DNA stampo). La reazione polimerasica, agendo ciclicamente (30 o più cicli), produrrà centinaia di milioni di copie del frammento di DNAds tra i due primer. Iniziando da una singola copia di DNA, con 32 cicli di PCR, si ottengono teoricamente più di un miliardo di copie (in pratica un poco meno), che corrispondono a circa una fmole. Utilizzando la PCR si usa anche il termine: <u>amplificare</u> (al posto di ricopiare o sintetizzare DNA) ed <u>amplificato</u> per il DNAds prodotto dalla PCR. Il DNA di interesse (es. di un intero gene) può essere amplificato da un DNAds molto più lungo (es. DNA di intero cromosoma) perché la lunghezza dell'amplificato è stabilita dalla posizione di associazione dei primer; inoltre la PCR amplifica specificamente il frammento di interesse anche se è in soluzione con milioni di altri frammenti di DNA.
La sequenza delle reazioni e dei cicli della tecnologia PCR sono descritti nella didascalia della figura 1-8
L'identità di sequenza del DNA amplificato con il DNA stampo è data dalla specificità dei primer utilizzati per la reazione DNA-polimerasica. L'associazionedei primer di circa 20 basi al DNA da amplificare è una ibridazione detta ad alta stringenza (riconoscimento molecolare specifico) come quella

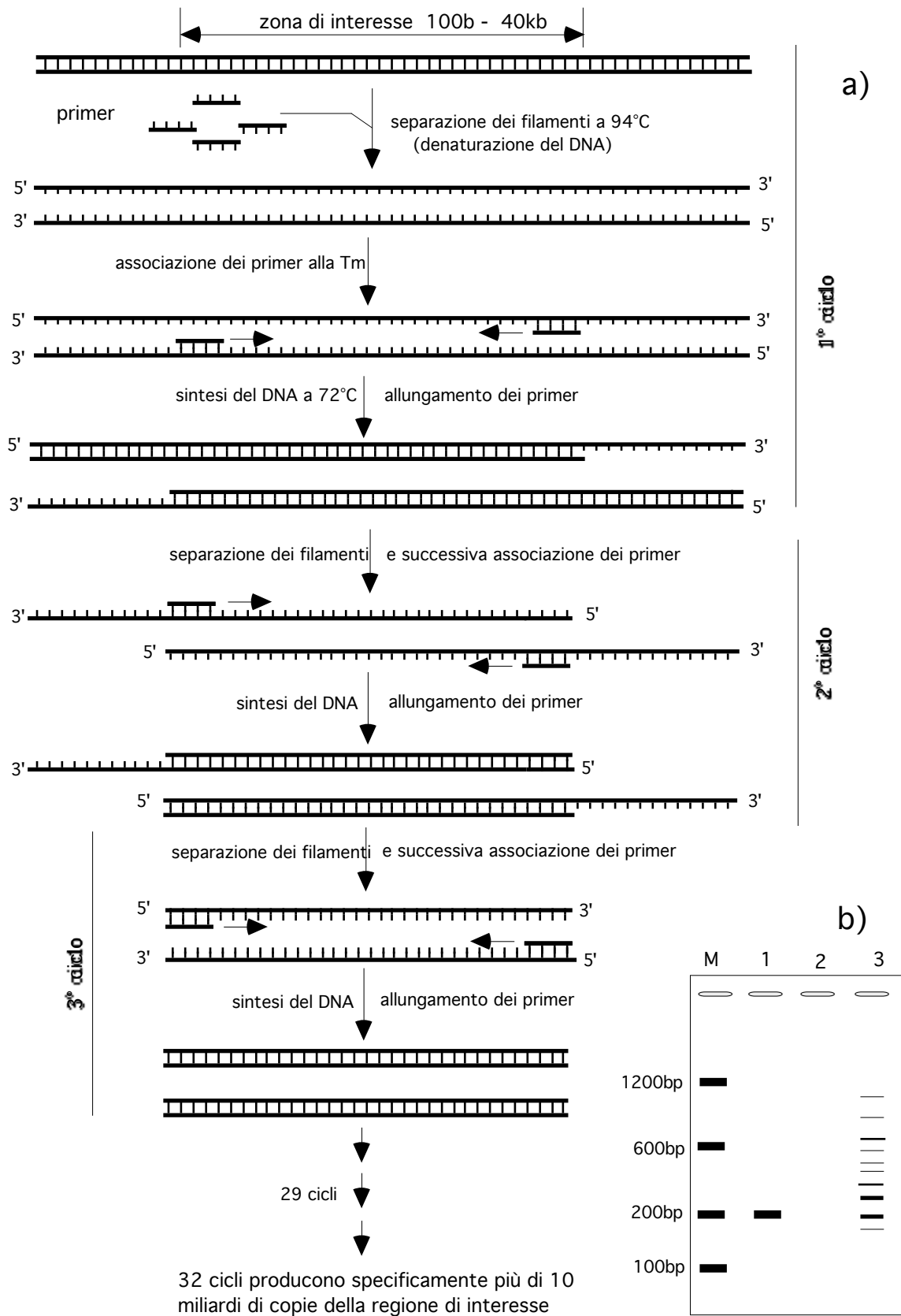



Figura 1-8. Tecnica della reazione a catena della DNA-polimerasi (PCR).

Figura 1-8a. Amplificazione di una piccola regione di DNA (200b) che in genere è utilizzata per analizzare la presenza di una data sequenza in campioni di DNA come nel vaglio delle genoteche. Con lo stesso meccanismo si possono amplificare regioni di decine di migliaia di basi.
La miscela di incubazione dall'inizio include tutti i componenti ed in quantità sufficienti per far procedere per alcune ore la reazione: i 4 dNTP, i due primer, DNA da ricopiare, ioni, tampone e l'enzima *Taq* polimerasi. La *Taq* polimerasi ha il suo ottimo di funzionamento a 72^0 C e non si denatura se esposta per tempi brevi a 94°C. Il volume di incubazione può essere molto piccolo ($5 \div 10\mu l$) quando il DNA da amplificare è molto scarso. La μprovetta che contiene l'incubazione è posta in una macchina (termociclizzatore) che può essere programmata come tempi, temperature e numero di cicli. Il ciclo standard programmato con la macchina è circa 1 minuto a 94° per denaturare il DNAds da amplificare, poi rapidamente la temperatura dell'incubazione è fatta scendere alla temperatura di associazione dei primer che è mantenuta costante per circa un altro minuto. La temperatura ottimale per l'associazione di ogni primer è calcolata tenendo conto della composizione della soluzione di sintesi e delle coppie G/C e A/T che si formeranno al momento dell'ibridazione. Questa temperatura è detta Tm ed è la T di fusione (melting) del complesso primer-DNA stampo. Si cerca di costruire i due primer con valori di Tm vicini (scegliendo opportunamente le sequenze e le lunghezze dei primer) e poi si opera alla T media delle Tm dei due primer che è in genere tra i 50-70°C.
Operare alla Tm riduce il numero di associazioni corrette tra primer e DNA-stampo ma evita che si formino associazioni aspecifiche che porterebbero ad amplificati non voluti. Se la T di associazione è troppo bassa i primer possono legarsi anche a sequenze non esattamente complementari, se è troppo alta si riduce la quantità di primer che si legano al DNA.
Dopo la fase di associazione si passa a quella di estensione dei primer, cioè di sintesi del DNA (circa 1 minuto a 72°) e qui termina il ciclo. Si procede nella stessa maniera per i successivi cicli: denaturazione-> associazione dei primer->estensione dei primer (mantenendo costante la T e i tempi delle tre fasi. Il tempo di denaturazione e di sintesi del DNA sono incrementati quando il DNA stampo è di notevoli dimensioni. Inoltre quando si vuole che i primer si associno più specificamente si stabiliscono T di associazione più alte delle Tm dei primer. Dopo 3 cicli, gli amplificati a doppio filamento risultano delle dimensioni programmate (zona di interesse delimitata dai primer). Al 5° ciclo, il 75% del DNAds amplificato ha le dimensioni programmate. Questi filamenti di DNAds a loro volta saranno utilizzati come stampo progressivamente sempre di più (a scapito di quelli di non esatta lunghezza) dato il loro progressivo e più rapido aumento di copie. Il numero dei filamenti di DNAds sintetizzati con la dimensione programmata aumenta così in progressione esponenziale: 2, 4, 8 e così avanti fino a compiere più di 30 cicli e produrre specificamente un miliardo di copie del DNA di interesse.

Figura 1-8b. Foto ai raggi UV (ultravioletti) del gel di elettroforesi di 3 amplificati di PCR. L’elettroforesi del campione 1 indica che la PCR si è svolta correttamente perché è stata sintetizzata una grande quantità di DNA delle dimensioni previste (200b) con l'utilizzo di 2 primer delimitanti una regione di interesse di 200b. L’elettroforesi del campione 2 indica che l’amplificazione non è avvenuta, perché il DNA stampo non includeva sequenze complementari ad uno o ambedue i filamenti. L’elettroforesi del campione 3 indica che l’amplificazione non si è svolta correttamente perché sono stati sintetizzati frammenti di DNA di varie dimensioni. In questo caso si assume che uno od ambedue i primer si siano associati a più di una regione del DNA stampo. Il gel contiene bromuro di etidio che esposto alla luce UV, rende visibile e fotografabile il DNA. M, marcatori di dimensione del DNAds (frammenti di DNAds dei quali si conosce il numero di b). P, pozzetti in cui sono stati posti i campioni. Altri dati nel testo.

dell'analisi Southern. Nella tecnica PCR il riconoscimento, essendo doppio, è più accurato: se un primer non si associa al DNA stampo la reazione DNA-polimerasica non procede.
Data la grande quantità di DNA che in genere viene prodotta, il DNA amplificato è visibile direttamente sul gel usato per l'elettroforesi aggiungendo alla soluzione di preparazione del gel di agarosio del bromuro di etidio, un colorante fluorescente. Questo composto ha un gruppo planare che si intercala tra le basi del DNAds e la vicinanza con le basi lo rende più fluorescente di quanto non lo sia quando è libero nel gel. Con lo stesso colorante si può colorare DNAss ed RNA (analisi Northern) ma la fluorescenza emessa è inferiore a quella del DNAds.
Dopo elettroforesi, esponendo il gel alla luce ultravioletta, si vedono le bande elettroforetiche, cioè l'ammasso di frammenti di DNA delle stesse dimensioni su una linea perpendicolare al flusso della corrente (figura 1-8b). Quando si analizza il prodotto della PCR si deve vedere una unica banda di forte intensità, contenente frammenti di DNA di lunghezza uguale a quella esistente tra i due estremi 5' dei due primer. Se si sono compiuti errori (nel definire la sequenza del DNA da amplificare o nella sintesi dei primer) i due primer non si legano al DNA stampo, la reazione della PCR non procede e sul gel non compare nessuna banda oltre a quella del DNA stampo (figura 1-8b). Talvolta, per associazione non corretta dei primer (es. a sequenze simili a quella di interesse), la PCR può produrre uno o più amplificati con sequenze diverse da quelle che dovevano essere ricopiate (figura 1-8b). Se esistono dubbi è necessario verificare l'identità del prodotto della PCR mediante determinazione della sequenza nucleotidica (figura 1-10) del DNA amplificato che deve essere identica al DNAds usato come stampo.

Altre applicazioni della reazione a catena della DNA-polimerasi

La tecnica della PCR è molto potente (può amplificare fino centinaia di milioni di copie partendo da due sole copie di una regione di DNA prelevato da una singola cellula) e versatile perché è utilizzata anche per analisi del DNA e del RNA, per vaglio di genoteche (figura 1-14), per la mappatura fisica (figura 3-11b), per determinare l'impronta ed il profilo del DNA (figure 3-15 e 3-16), l'analisi dei marcatori dei geni (figure 3-8 e 3-9), per realizzare mutazioni sito specifiche (figura 4-3) e utilizzando un singolo primer, per determinare la sequenze di DNA (figura 1-10).

Reazione della trascrittasi inversa accoppiata alla reazione a catena della DNA-polimerasi (RT-PCR). Gli mRNA estratti da un tessuto (mRNA totale) sono convertiti nei rispettivi cDNA mediante l'enzima trascrittasi inversa (figura 1-11). Utilizzando specifici primer con la PCR si amplifica esclusivamente il cDNA di interesse per rivelarne la presenza e/o per averne sufficienti quantità per analisi o per transfezione ed espressione.

La tecnologia della PCR analitica.

La tecnologia della PCR è utilizzata anche per l'analisi sequenza-specifica del DNA. I primer operano simultaneamente due ibridazioni su uno stesso frammento di DNAds, ed avendo una dimensione di 20b possono ibridare specificamente su sequenze uniche del genoma umano e quindi individuare il frammento di DNA di interesse tra milioni di frammenti di DNA o, direttamente sul DNA genomico non digerito, individuare una regione cromosomica avente la sequenza nucleotidica identica alla sequenza di interesse. La specificità della doppia ibridazione è mostrata dalla sintesi di ogni singola copia del DNAds di interesse (figura 1-8). La PCR amplificando il DNA di interesse ha una altissima sensibilità (capacità di rivelare piccolissime quantità di DNA) che è circa 10.000 volte superiore a quella dell'analisi Southern (ibridazione non seguita da amplificazione). La proporzione è tale che con la tecnica PCR si riesce a vedere le quantità pari ad un colonnina (circa un metro) del Prato dei Miracoli di Pisa, con l'analisi Southern si riesce a rivelare solo quantità superiori al monte Everest.

La tecnica della PCR è preferita all'analisi Southern anche per la semplicità di esecuzione e perché evita l'uso di isotopi radioattivi.

La PCR analitica è identica a quella utilizzata per amplificare il DNA (figura 1-8), per essa non è necessario amplificare tutto il DNA di interesse ma solo una parte specifica di esso che può essere anche solo di 100b.

Si sottopone l'insieme dei frammenti diversi di DNA all'amplificazione con PCR e la presenza del DNA di interesse è verificata analizzando l'amplificato su gel di agarosio (figura 1-8b). La presenza di una unica banda delle dimensioni attese dà l'indicazione che nella mistura dei vari frammenti di DNA è presente un frammento capace di ibridare con i due primer. Successivamente, se necessario, si può confermare la risposta positiva avuta con la PCR identificando l'amplificato mediante determinazione della sua sequenza nucleotidica.

La PCR è anche utilizzata per rivelare infezioni virali e batteriche (amplificando i DNA o gli RNA degli agenti patogeni presenti anche in piccolissime quantità nei liquidi biologici o tessuti), inoltre con la PCR può identificare mutazioni cancerogene specifiche nel DNA genomico di una cellula cancerosa tra 10^6 cellule normali.

PCR analitica quantitativa (Real-time PCR).

La tecnica della PCR quantitativa (Real-time PCR) è utilizzata per determinare la concentrazione specificamente di un frammento di DNA in una soluzione contenente milioni di altri frammenti di DNA. Essa permette di monitorare, in tempo reale, la quantità di DNA amplificato mentre procede la reazione della PCR e, dall'andamento dei valori di DNA amplificato nel tempo (misurato come numero di cicli), determinare, con alta approssimazione, la quantità del DNA stampo di interesse.

Per osservare l'andamento della reazione della PCR viene aggiunto alla miscela di reazione fluorocromo, come il colorante fluorescente "verde-SYBR I", che si lega

specificamente al DNAds intercalandosi tra le basi e, quando eccitato, è capace di emettere una fluorescenza molto intensa. Gli incrementi di fluorescenza sono registrati da apparecchiature ottiche, elettroniche e grafiche, annesse al termociclizzatore, che permettono di seguire il procedere della reazione della PCR e di registrare su carta il grafico dell'intensità della fluorescenza in funzione del numero di cicli della PCR (figura 1-9).

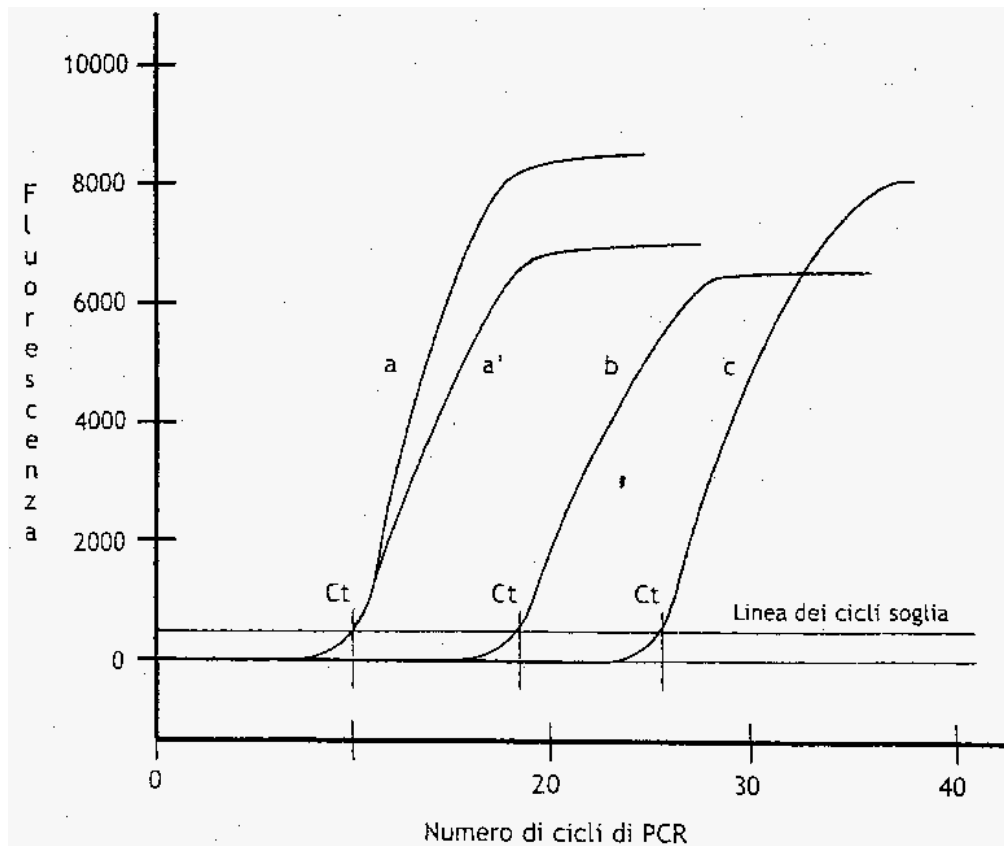



Figura 1-9. Grafico della fluorescenza emessa durante l'amplificazione del DNA stampo in funzione dei cicli di PCR utilizzando il fluorocromo Verde SYBR. a, a', b, c, sono le curve di amplificazione di uno stesso di DNA aggiunto in concentrazioni: [a] > [a'] e [a] > [b] > [c]. La linea soglia dei cicli unisce i punti detti soglia del ciclo (Ct), numero di cicli in cui le curve di amplificazione di uno stesso DNA stampo assumono andamento lineare. Il grafico mostra che il numero dei cicli soglia è inversamente proporzionale alla concentrazione iniziale di uno stesso DNA stampo e che, al termine della PCR, la quantità totale di DNA stampo amplificato non è in relazione alla quantità iniziale del DNA stampo.

Durante i primi cicli della PCR la quantità di DNAds sintetizzato è sotto il valore registrabile dalla strumentazione. Sul grafico appare una linea pressoché orizzontale poi dopo un certo numero di cicli (12-15 cicli), la curva si impenna verso l'alto e procede con andamento lineare. Quel punto della curva corrispondente al numero di cicli in cui l'amplificazione del DNA stampo è divenuta logaritmica. Se in una miscela di reazione della PCR è presente, rispetto ad un'altra, una quantità doppia di DNA stampo, la fase di amplificazione

logaritmica della PCR inizierà più precocemente e ciò sarà mostrato da una curva che si flette verso l'alto ad un numero inferiore di cicli della PCR.
Si dimostra così che il numero di cicli, in cui la curva inizia ad assumere l'andamento lineare, è inversamente proporzionale alla concentrazione del DNA stampo, presente nella miscela di reazione della PCR. Pertanto da quel numero di cicli si può risalire specificamente alle quantità ignote di DNA stampo sottoposte alla PCR quantitativa, e data la specificità dei primer per il DNA di interesse ciò avviene anche quando nella miscela di reazione della PCR sono presenti altre specie molecolari di DNA. Il punto in cui la curva assume l'andamento lineare è detto soglia del ciclo (Ct) e corrisponde ad un dato numero di cicli per ogni diversa concentrazione di uno stesso DNA stampo. I Ct di uno stesso DNA stampo presente in concentrazioni diverse nel grafico sono allineati su una retta orizzontale detta linea di soglia dei cicli (figura 1-9).
Le quantità assolute (nanogrammi) del DNA amplificato sono stabilite confrontando l’intensità della fluorescenza ottenuta con la PCR quantitativa del DNA di interesse con l’intensità della fluorescenza ottenuta con quantità note di DNA (quantità standard) nelle stesse condizioni di PCR quantitativa.
Ripetendo più analisi di PCR quantitativa con la stessa quantità di uno stesso DNA stampo si osserva che la fase di amplificazione logaritmica inizia allo stesso numero di cicli, poi, dopo la fase logaritmica, l'amplificazione rallenta e cessa a valori diversi di fluorescenza massima. Ciò è mostrato dalle curve che passano dalla fase lineare a quella di plateau. I diversi livelli di fluorescenza dei plateau indicano che, al termine della PCR, sono state sintetizzate quantità diverse di DNA (cioè nelle varie incubazioni lo stesso DNA stampo è stato copiato un numero diverso di volte). La causa di queste differenze di quantità di sintesi di DNA sono attribuite all'esaurimento dei substrati e alla riassociazione dei filamenti di DNA replicato. Data la loro alta concentrazione i filamenti sintetizzati tendono a riassociarsi tra loro piuttosto che associarsi ai primer che, con il procedere dei cicli, diminuiscono in concentrazione. Tuttavia non si spiega perché queste cause provochino effetti diversi sulla quantità di DNA sintetizzato in identiche condizioni di partenza di PCR, ed in particolare con le stesse quantità di reagenti.
Le prove della PCR quantitativa ripetute con la stessa quantità di DNA stampo dimostrano che registrando il numero di cicli in cui la reazione della PCR inizia la fase logaritmica si ha l'indicazione corretta della quantità del DNA stampo, mentre valutando la quantità del DNA sintetizzato alla fine della reazione (come si fa con la PCR per scopi sintetici) non si può avere una misura quantitativa del DNA stampo amplificato.
La PCR quantitativa è utilizzata nella sperimentazione e nelle analisi cliniche per analizzare quantitativamente l'espressione di singoli geni misurando la concentrazione del relativo mRNA presente nelle cellule. Lo mRNA totale, estratto dalle cellule, è convertito in cDNA mediante RT-PCR e quindi è sottoposto alla PCR quantitativa specifica per il cDNA di interesse. La PCR quantitativa è utilizzata anche per controllare le quantità di ibridi DNA-RNA o DNA-cDNA che si sono formati nei microarray (capitolo 2).

La PCR quantitativa che utilizza il verde SYBR ed una curva standard è uno dei metodi di PCR quantitativa. Altri metodi utilizzano tipi diversi di sonde fluorescenti sequenza-specifiche.

La tecnica PCR analitica quantitativa ha progressivamente sostituito le analisi Southern e Northern.
Inizialmente la tecnica PCR aveva sostituito le analisi Southern e Northern per la sua alta sensibilità e maggiore semplicità di esecuzione, mentre le analisi Southern e Northern continuavano ad essere utilizzate per le analisi quantitative del DNA ed RNA. Negli anni '90 con l'avvento della PCR quantitativa, le tecnologie Southern e Northern sono state sostituite con questa nuova tecnica che è più accurata (se ripetuta riproduce gli stessi valori che sono più vicini alla realtà) ed è capace di analizzare quantità di DNA o RNA molto inferiori a quelle richieste dalle analisi Southern e Northern. L'analisi Southern è ancora utilizzata quando si ricerca DNA del quale non è nota l'esatta sequenza. Ad esempio quando si utilizza una sonda specifica del gene di una specie per ricercare il gene omologo di un'altra specie, sapendo che la sequenza dei due geni è simile ma non identica. In queste condizioni è più probabile associare una sonda che due primer per la PCR. E' ancora necessario conoscere le analisi Southern e Northern perché molti ed importanti dati sperimentali riportati nei testi di biologia molecolare sono stati ottenuti con queste tecniche.

Errori durante la sintesi *in vivo* ed *in vitro* del DNA
Alcune considerazioni sulla sintesi del DNA mediante le tecniche di subclonazione e PCR.

L'alta fedeltà riscontrata nella replicazione del DNA nelle cellule batteriche ed eucariotiche, risulta da una bassa frequenza di errori durante la sintesi del DNA e da una correzione degli errori commessi durante la stessa sintesi. Le DNA-polimerasi, oltre all'attività sintetica hanno anche un'attività esonucleasica 3'->5', che procede in senso opposto a quella di sintesi. Quando casualmente viene inserito nella catena nascente di DNA un nucleotide errato, il mancato accoppiamento delle basi determina un allargamento locale della doppia elica che è percepito dall'enzima DNA-polimerasi. L'enzima blocca la sua attività sintetica, inizia a procedere in senso inverso ed attiva l'attività esonucleasica, eliminando, uno per volta, il nucleotide errato ed altri 2-3 nucleotidi; quando il DNA ha un corretto accoppiamento di basi l'enzima riprende la sintesi del DNA.
Nucleotidi errati che sfuggono a questo controllo sono eliminati (salvo rarissime eccezioni) dai sistemi di riparazione del DNA (appendice C).
La sintesi del DNA per subclonazione, essendo operata all'interno di cellule batteriche, riproduce fedelmente il DNA del costrutto plasmide-DNA-inserto di interesse.
Quando il DNA è replicato *in vitro* mediante PCR catalizzata dall'enzima Taq-polimerasi, che non possiede attività esonucleasica 3'->5', la frequenza di errore durante la sintesi del DNA è alta.

Un frammento di DNA di 1kb, usato come stampo, che sottostà a 20 cicli di PCR viene amplificato fino a dare circa 260.000 copie. Circa il 40% dei nuovi frammenti di DNA contiene una base errata (ed alcuni di essi due) conseguente all'assenza dei meccanismi di correzione e di riparazione. Tutto ciò risulta in una miscela di frammenti aventi lo stesso numero di basi con sequenze molto simili ma non identiche per una o due basi, anche se il DNA stampo utilizzato era composto di frammenti di DNA aventi tutti la stessa identica sequenza. Poiché l'inserimento di basi errate avviene a caso, la posizione della base errata risulta diversa nei diversi filamenti e ciò permette di ottenere comunque una sequenza corretta mediante le tecnologie standard per la determinazione della sequenza nucleotidica (figura 1-10). Per ciascuna posizione delle basi (1a base, 2a base, 3a base, ecc.) il numero di basi errate nei frammenti di DNA amplificato risulta irrilevante rispetto al numero di basi corrette dei rimanenti frammenti (questi frammenti includono sia i frammenti che non hanno nessuna base errata che i frammenti che hanno una base errata in una posizione diversa da quella presa in esame). Su 260.000 copie di DNA, 156.000 hanno la sequenza corretta e 104.000 hanno una base errata che può essere in una delle 1000 posizioni del DNA stampo. Quindi in ciascuna posizione della sequenza si potrà avere una base e molto raramente due basi errate contro le 259.000 o 258.000 corrette. Pertanto le tecnologie di sequenziamento (manuali o automatiche) analizzando l'amplificato dalla PCR forniranno la sequenza corretta del DNA stampo perché la loro sensibilità non è influenzata da una o due basi diverse in presenza di una grande maggioranza di basi corrette presenti nella stessa posizione di sequenza. Se sul grafico il picco della base corretta è 1 cm, il picco della base errata (sovrapposto a quello della base corretta) è meno di 1/15 di millimetro, in genere non visibile.
La frequenza di errore della PCR può essere ridotta utilizzando la DNA-polimerasi di *Pyrococcus furiosus* che è termostabile e possiede attività esonucleasica 3->5'. Con questo enzima un DNA stampo di 1kb, sottoposto a 20 cicli di PCR risulta avere una base errata nel circa 3,5% dei frammenti amplificati (cioè in 9100 copie delle 260.000 amplificate, un errore ogni circa 28.570 basi ricopiate).
Il DNA amplificato con PCR può essere utilizzato per determinare la sequenza del DNA stampo, mentre sorgono problemi se il DNA amplificato deve essere subclonato e successivamente analizzata l'attività molecolare o la funzione fisiologica del suo prodotto. Ad esempio: il dosaggio dell'attività molecolare della proteina espressa da un cDNA subclonato. Quando un cDNA, amplificato con PCR, viene inserito in un plasmide e successivamente si opera la subclonazione, può accadere (sfortunatamente) di avere isolato proprio un clone contenente un cDNA inserto con una base mutata durante la sua sintesi con la PCR. A causa di questa base errata la proteina espressa può risultare inattiva. Il problema è ovviato operando la selezione di più cloni, determinando la sequenza del loro cDNA inserto e scegliendo un clone avente l'inserto cDNA con la sequenza corretta, identica a quella del cDNA usato come stampo per la

PCR. Quel clone sarà utilizzato per sintetizzare *in vitro* la relativa proteina per poi effettuare su essa le prove funzionali.
Può sembrare una perdita di tempo usare prima la PCR invece di clonare subito il cDNA. Tuttavia con la PCR, scegliendo opportunamente i primer, si può amplificare esclusivamente la regione del DNA di interesse e poi aggiungere adattatori (come visto in figura 1-11 o con primer includenti un sito di restrizione figura 22-a) per operare successivamente la subclonazione. Con la subclonazione occorre operare utilizzando i siti di restrizione che invece sono punti obbligati della sequenza e la loro posizione può creare problemi (troppo distanti/troppo vicini o all'interno dalla regione di interesse).
Con la subclonazione l'insorgere di una mutazione per errore di sintesi è un evento raro ma sempre possibile per cui un controllo della sequenza del DNA di interesse subclonato viene sempre eseguito.

Tecnologia per la determinazione della sequenza nucleotidica del DNA secondo il metodo di Sanger

Ogni nuovo frammento di DNA, subclonato (figura 1-7) o clonato (figura 1-14), è sottoposto all'analisi della sua sequenza nucleotidica.
La sequenza nucleotidica caratterizza specificamente la molecola di DNA come la formula di struttura identifica le molecole dei composti organici, inoltre in essa è codificata la memoria chimica delle caratteristiche genetiche degli esseri viventi.
La tecnologia di Sanger essendo l'unica tecnica capace di determinare la sequenza nucleotidica, è anche l'unica capace di identificare inequivocabilmente un frammento di DNA ancora ignoto.
Le altre tecnologie di identificazione di frammenti di DNA, ibridazione, analisi Southern e PCR sono capaci di identificare specificamente un frammento di DNA tra milioni di altri solo se esse sono state adattate ad operare alla Tm della sequenza del frammento di interesse. Sequenza che necessariamente deve essere stata precedentemente determinata con la tecnica di Sanger per stabilire la Tm dell'ibrido di interesse.
Frederick Sanger ha vinto due volte il premio Nobel per la chimica: nel 1958 da solo, per studi sulle proteine, in particolare per aver definito la sequenza aminoacidica dell'insulina e nel 1980 insieme a Walter Gilbert che aveva inventato un diverso metodo di analisi della sequenza del DNA e a Paul Berg per i suoi (di Berg) studi sul DNA ricombinante. Il metodo di analisi proposto da Gilbert è stato abbandonato perché troppo laborioso mentre quello di Sanger è usato nei laboratori di tutto il mondo.
La tecnologia per determinare la sequenza nucleotidica del DNA secondo il metodo di Sanger (figura 1-10) utilizza 2',3'dideossinucleotidi trifosfati (ddNTP, detti anche terminators = terminatori della sintesi del DNA) per arrestare la reazione di sintesi del DNA specificamente a livello delle quattro basi azotate. In quattro microprovette, vengono aggiunte quattro soluzioni contenenti, il primer

(un oligonucleotide di circa 20 basi complementare alla regione al 3' del DNA da analizzare), l'enzima DNA-polimerasi, i dNTP ed un dNTP radioattivo (in genere ATP*) ed un ddNTP diverso (ddCTP, ddATP, ddTTP, ddGTP) in ciascuna incubazione. Le 4 soluzioni di reazione vengono portate a circa 30°C e viene aggiunto il DNAds stampo (precedentemente denaturato a circa 65°C e poi raffreddato lentamente a 30°C). Nella soluzione di incubazione ora il primer è associato al filamento stampo di DNAss (denaturato) e l'enzima DNA-polimerasi inizia a sintetizzare DNA. Quando un ddNTP (es. il ddCTP in figura 1-10a) viene incorporato nel filamento di DNA nascente, esso non permette un'ulteriore reazione con un dNTP perché i ddNTP mancano dell'ossidrile al 3'. La quantità dei ddNTP è stabilità in quantità molto inferiore (circa 1/10) rispetto a quella dei dNTP, pertanto l'inserimento casuale del ddCTP è meno frequente dell'inserimento casuale del dCTP e ciò permette di sintetizzare tanti filamenti di lunghezza diversa quante sono le basi G contenute nel DNA stampo. Se la normale miscela di reazione contenesse solo ddCTP la sintesi si arresterebbe alla prima base G del DNA stampo, mentre in presenza di soli dCTP si formerebbe solo un filamento lungo quanto il DNA stampo. Con ddCTP in concentrazione pari ad un decimo di quella di dCTP, dopo 2-5 minuti verranno sintetizzate molte copie di vari filamenti aventi numero di basi diverse, tutti con al loro estremo 3' un ddNTP. Ogni specie di filamento differirà rispetto all'altro in lunghezza (numero di basi) e la differenza minima tra un filamento e l'altro sarà di una singola base.
I filamenti neosintetizzati vengono frazionati mediante un elettroforesi su gel di poliacrilamide. L'acrilamide polimerizzando forma un gel (poliacrilamide) che ha pori più piccoli di quelli del gel di agarosio e permette di separare filamenti di DNAss di lunghezza diversa anche di una singola base. Al termine della reazione di sintesi del DNA, le incubazioni sono scaldate per circa un minuto a 95°C per denaturare il DNA, poi sono poste sul gel di poliacrilamide contenente urea (agente denaturante) e l'elettroforesi è operata ad un voltaggio tale da garantire un passaggio di corrente che faccia migrare i filamenti di DNAss e mantenga la temperatura del gel a circa 55°C per evitare la riassociazione dei filamenti di DNA. Il mantenimento delle condizioni denaturanti il DNA è necessario affinché i filamenti di DNA neosintetizzato possano essere frazionati durante l'elettroforesi in base alla loro lunghezza. Se il DNA rimanesse a doppio filamento (DNA-stampo e DNA-neosintetizzato di varia lunghezza) sarebbe difficile osservare le differenze di lunghezza dei filamenti neosintetizzati. I filamenti neosintetizzati di DNAss sono radioattivi per aver incorporato ATP* radioattivo ed essendo carichi negativamente migreranno verso il polo positivo. Al termine dell'elettroforesi, i filamenti di DNAss neosintetizzati sono disposti dal più lungo al più corto ordinatamente dal polo negativo verso quello positivo (parte bassa del gel). Il DNA stampo migrerà in un'unica banda nella parte alta del gel e risulterà invisibile nell'autoradiografia perché non radioattivo.
Il gel viene disidratato su un foglio di carta da filtro ed assume la consistenza di una sottile lamina di plastica non appiccicosa. Su esso viene posto un film radiografico ed effettuata l'autoradiografia. I filamenti di DNA neosintetizzato

hanno lunghezze diverse in relazione al ddNTP presente nell'incubazione in cui sono stati sintetizzati ed in relazione alla posizione nella sequenza del DNA stampo della base complementare allo stesso ddNTP. In figura 1-10, i numeri (1, 2, 3, 4) indicano i filamenti sintetizzati in presenza di ddCTP e le corrispondenti bande autoradiografiche (la banda include molte copie dello stesso filamento di DNAss sintetizzato).
La sequenza è dedotta leggendo la base complementare al ddNTP presente nell'incubazione iniziando a leggere nell'autoradiografia dalla banda più vicina al polo positivo perché i filamenti neosintetizzati più brevi (vicini al polo positivo) identificano le basi più vicine al 3' del frammento di DNA sequenziato. In figura 1-10a è descritta solo la reazione di sintesi con il ddCTP, le altre reazioni con ddATP, ddGTP e ddTTP sono eseguite nello stesso modo e contemporaneamente ad essa.

La lunghezza del filamento (indicata dalla posizione della banda nell'autoradiografia) stabilisce la posizione della base nella sequenza del DNA stampo ed il ddNTP presente nella soluzione di reazione indica il tipo di base (A, T, G, o C) ad esso complementare che occupa quella posizione.

In passato, per determinare la sequenza nucleotidica si usava purificare un filamento di DNAss del DNAds di interesse, successivamente la procedura di purificazione è stata abbandonata quando si è visto che la stessa tecnica di Sanger poteva determinare la sequenza di frammenti di DNAds.
Quando il frammento del DNAds del quale si è determinata la sequenza non è omogeneo, cioè non è un'unica specie molecolare ma ad esempio due, le basi diverse presenti nei due frammenti sono rivelate dalla presenza nell'autoradiografia di due bande della stessa intensità, alla stessa distanza dall'origine dell'elettroforesi ma in due corsie diverse della stessa elettroforesi (figura 1-10c).
Soluzioni contenenti frammenti di DNA diversi, solo per una singola base, si ottengono quando il DNA di interesse è estratto da cellule diploidi come le cellule somatiche umane. Il DNA estratto è digerito con enzimi di restrizione che taglieranno il DNA di interesse da ambedue gli alleli e poi, avendo i due frammenti di DNA la stessa lunghezza, saranno purificati insieme (non separati) mediante elettroforesi. Se si usa la tecnica PCR, si ottiene un amplificato contenente anch'esso due frammenti di ugual lunghezza, ma diversi per una base, perché i primer si assoceranno alle stesse sequenze di DNA dei due alleli producendo le due specie di amplificati. Per avere soluzioni di frammenti di DNA omogenei (frammenti aventi tutti la stessa sequenza) occorre clonare il DNA digerito e purificato con elettroforesi o amplificato mediante PCR.

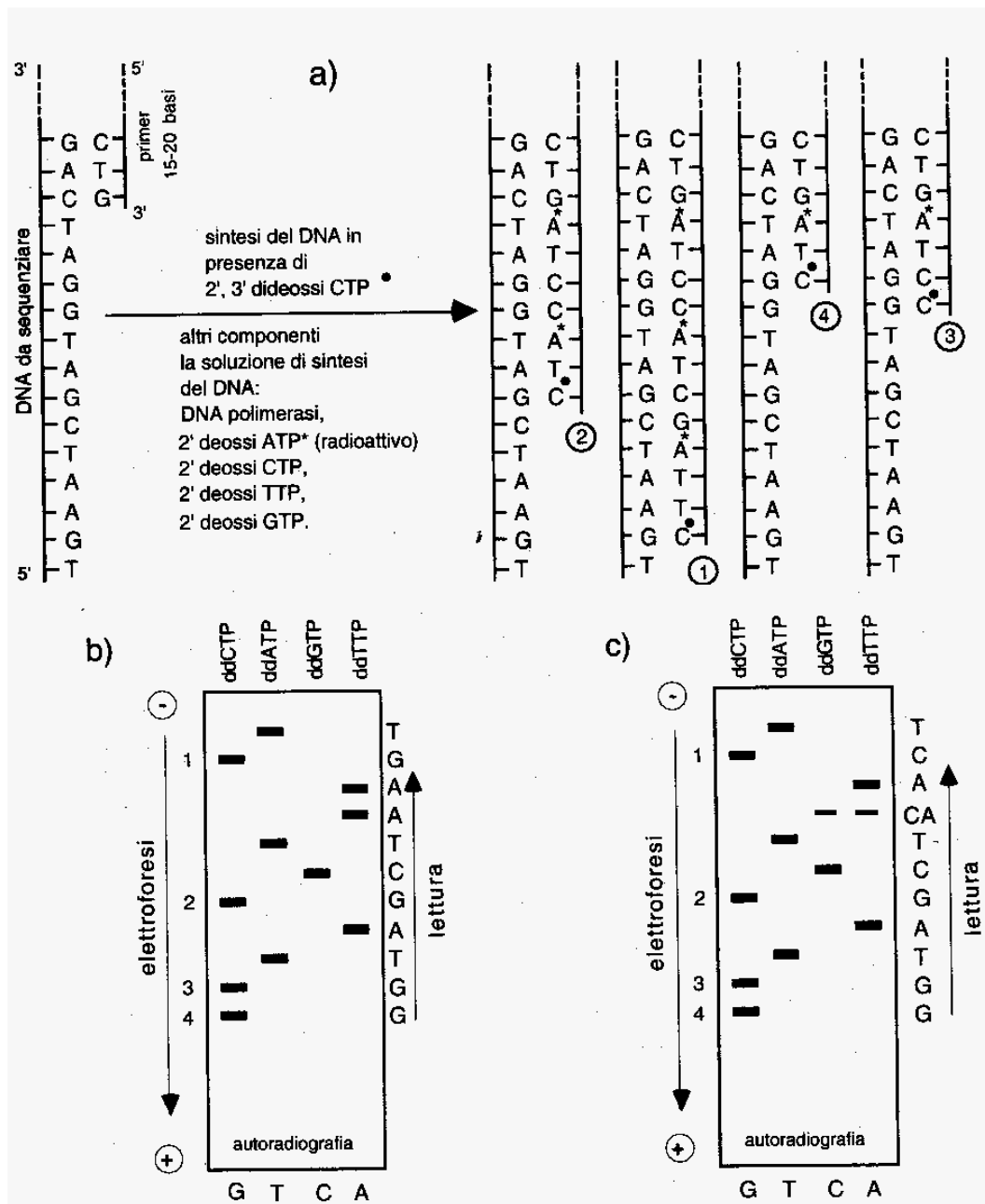



Figura 1-10. Schema della tecnica per determinare la sequenza nucleotidica del DNA secondo il metodo manuale di Sanger.

a). Schema della reazione di sintesi del DNA in presenza del ddCTP. Le altre reazioni rispettivamente in presenza di ddATP, ddGTP e ddTTP sono eseguite in parallelo (non sono indicate).

b). Autoradiografia eseguita sul gel di poliacrilamide disseccato sul quale, mediante elettroforesi denaturante, sono stati separati i frammenti di DNAss neosintetizzati. I ddNTP usati nell'incubazione sono indicati in alto dell'autoradiografia e le basi ad esso complementari in basso.

c). Autoradiografia eseguita sul gel di poliacrilamide disseccato sul quale, mediante elettroforesi denaturante, sono stati separati i frammenti di DNAss neosintetizzati da un DNAds stampo non omogeneo, contenente 2 specie di frammenti di DNAds, presenti in eguali quantità e diversi per 1 base (A-C) nella posizione N° 8 dal basso. Questa sequenza indica che i due alleli sono eterozigoti

Quando è nota la sequenza di almeno una parte del frammento di DNA da analizzare, si utilizza un primer di circa 20b sintetizzato *in vitro* con la sequenza progettata identica alla parte terminale al 5' del frammento di interesse (come in figura 1-10). Se la sequenza del frammento è completamente ignota si può usare un oligonucleotide sintetico avente la sequenza complementare a quella del vettore (che è nota) con il quale si è clonato il frammento di interesse. Si inizia a determinare la sequenza del costrutto continuando fino a quando la sequenza del vettore termina ed inizia quella del frammento di interesse. Questo tipo di primer è detto <u>universale</u> perché può essere utilizzato per determinare la sequenza di qualsiasi frammento clonato con quel vettore.
Una analisi di sequenza rivela 200-500 basi consecutive, pertanto se il frammento è lungo più di questa misura, dopo la prima determinazione occorre sintetizzare un nuovo primer avente la sequenza identica alla parte terminale al 3' della sequenza neodeterminata. E così avanti fino a quando si è determinata tutta la sequenza del frammento di interesse. Questo modo di procedere è detto "camminare sul DNA" e, se il DNA è del genoma, "camminare sul cromosoma" (chromosome walking)(figura 4-2).
Quando occorre determinare la sequenza di un frammento di DNA di sequenza completamente ignota e privo di vettore si può usare una miscela di oligonucleotidi aventi sequenze diverse "primer a caso" (random primer). Se uno di questi si lega al frammento di interesse, assumiamo a metà di esso, si determina solo una parte della sequenza. Tuttavia successivamente si può sintetizzare un primer complementare alla sequenza già determinata e determinare la parte di sequenza mancante. Utilizzando la regola di Chargaff si costruisce la sequenza completa del DNA stampo.
Nel tempo il metodo di Sanger (metodo che usa i ddNTP) è stato semplificato da una serie di geniali modifiche che hanno portato ad eseguire la reazione di sintesi del DNA in una unica incubazione, ad analizzare l'elettroforetogramma con macchine dotate di congegni ottici di alta precisione.
La più importante modifica è stata la legatura covalente di una diversa sostanza fluorescente a ciascuno dei quattro dideossi-NTP (ddCTP, ddATP, ddGTP, ddTTP). Ciascuna di queste sostanze quando è colpita da radiazioni elettromagnetiche emette una radiazione di diversa lunghezza d'onda che permette di identificare il ddNTP ad essa legato. Ciò ha permesso di effettuare la sintesi del DNA in una unica incubazione (invece di quattro). Dopo elettroforesi dell'unica incubazione, la sequenza dei nucleotidi è individuata in una unica corsia elettroforetica sulla base della migrazione delle bande nell'elettroforesi (tante quante sono le basi analizzate) e dalla lunghezza d'onda alla quale, opportunamente sollecitata, emette la sostanza fluorescente legata al ddNTP.
La fluorescenza ad una diversa lunghezza d'onda permette di identificare ciascuno dei quattro ddNTP e quindi di effettuare la sintesi del DNA in una unica soluzione e di frazionare i filamenti neosintetizzati in una unica corsia di elettroforesi.

I filamenti stampo e quelli sintetizzati che non hanno incorporato il ddNTP, non essendo fluorescenti, non sono registrati dall'apparecchiatura ottica.
Con i 4 ddNTP diversamente fluorescenti, si può operare la sintesi dei filamenti di DNA utilizzando il termociclizzatore per la PRC con una sequenza di cicli simile a quella per amplificare il DNA (figura 1-8) ma utilizzando un solo primer, perché si vuole analizzare un solo filamento del DNAds di interesse.
I filamenti sintetizzati aumenteranno in maniera lineare (e non esponenziale come avviene nella PCR classica), perché mancando il secondo primer, viene copiato un solo filamento del DNA stampo.
La tecnologia di analisi della sequenza con la PCR è detta "determinazione ciclica della sequenza nucleotidica" o "determinazione ad amplificazione lineare della sequenza nucleotidica" e consiste in una normale mistura di incubazione per determinare la sequenza del DNA: un primer, 4 dNTP e i 4 ddNTP diversamente fluorescenti. I cicli a T diverse sono come nella PCR: denaturazione--> associazione dei primer-->sintesi del DNA. La sintesi dei filamenti verrà bloccata dall'inserimento casuale dei diversi ddNTP. Lo stesso ciclo viene ripetuto molte volte al fine di sintetizzare più copie dei vari filamenti diversi per numero di basi. Pertanto la tecnologia della determinazione ciclica della sequenza nucleotidica permette di determinare la sequenza del DNA utilizzando quantità di DNA stampo molto inferiori rispetto a quelle necessarie per una analisi non ciclica del DNA che utilizza una sola volta il DNA stampo.
Utilizzando la metodologia dei quattro ddNTP diversamente fluorescenti è stato costruito un analizzatore automatico di sequenze nucleotidiche corredato di una apparecchiatura ottica che analizza la singola traccia elettroforetica con un sottile raggio laser e determina la sequenza in base alla posizione nell'elettroforetogramma di ciascuna banda e individua il tipo di base leggendo la lunghezza d'onda emessa per fluorescenza dal ddNTP del filamento di DNAss presente in varie copie nella stessa banda elettroforetica. I segnali raccolti dall'apparecchiatura ottica sono trasmessi ad un computer che li trasforma e presenta sul monitor, come un grafico continuo di picchi, ordinati secondo la sequenza analizzata e di colore diverso in relazione al tipo di base. Sotto ciascun picco è indicata la base corrispondente al colore. L'altezza dei picchi indica la quantità delle basi analizzate. Nel caso che il DNA analizzato provenga da alleli diversi per una base, in quella stessa posizione ci saranno due picchi coincidenti di colore diverso, di circa la stessa altezza ma di minore altezza rispetto ai picchi delle altre posizioni della sequenza.

Nota: con le attuali tecnologie per determinare le sequenze nucleotidiche si opera sempre su filamenti di DNAds (genomico o cDNA). Per convenzione anche se la sequenza determinata è di un cDNA (quindi di un mRNA), non si sostituisce la base T con la base U, cioè si scrive e si conserva nelle banche dati la sequenza del filamento senso del cDNA e non quella del mRNA da cui è stato sintetizzato il cDNA.

L'invenzione della tecnologia per la determinazione della sequenza è una grande conquista della biologia molecolare perché la sequenza nucleotidica

caratterizza specificamente un frammento di DNA, non solo chimicamente come polimero eterologo, ma anche perché svela l'informazione genetica in esso contenuta. La natura diversa dei geni è nella loro sequenza dato che la loro reattività chimica è pressoché identica, tuttavia con l'intervento delle proteine che riconoscono i diversi tratti di sequenza nucleotidica, la reattività data dalle interazioni deboli tra DNA e proteine viene utilizzata per realizzare l'espressione genica.
L'analisi della sequenza (data la riduzione dei tempi e dei costi della determinazione automatica delle sequenze nucleotidiche) è divenuta il miglior metodo per identificare un frammento di DNA proprio perché rivela la sequenza nucleotidica del frammento utilizzando la formazione di legami covalenti che sono stabili alle temperature sufficientemente alte per denaturare il DNA (95°C).
Le altre tecnologie di analisi del DNA, PCR ed ibridazione possono essere utilizzate per identificare un frammento di DNA solo se di questo frammento è stata precedentemente determinata la sequenza con il metodo di Sanger. Infatti solo conoscendo la sequenza del DNA si possono sintetizzare i primer per la PCR e le sonde per l'ibridazione, stabilire le Tm per l'associazione dei primer e per l'ibridazione delle sonde. Senza queste conoscenze la PCR e l'ibridazione con sonde non potrebbero essere tecniche per identificare frammenti di DNA.
Inoltre la conoscenza della sequenza permette la manipolazione specifica dello stesso frammento (ricombinazione, espressione in vitro dei cDNA e le tecnologie per lo studio della funzione dei geni, capitolo 2).
Un'altra particolarità del metodo di Sanger è quella di essere una analisi che opera mediante una sintesi covalente come abbiamo visto per l'analisi operate con la tecnica della PCR analitica.

Tecnologie per la costruzione delle genoteche

Le genoteche sono di due tipi: genoteche genomiche, costituite da DNA dei cromosomi nucleari e genoteche di cDNA (DNA complementare). Il cDNA è DNA a doppio filamento (DNAds), un filamento del quale è stato sintetizzato ricopiando un mRNA con l'enzima trascrittasi inversa. Poi, con altri enzimi, viene sintetizzato l'altro filamento di DNA che avrà la sequenza identica a quella del mRNA. Alle singole molecole di cDNA così sintetizzate vengono poi aggiunti degli adattatori contenenti una sequenza di restrizione per poter ricombinare *in vitro* il cDNA (figura 1-11). La conversione del mRNA in cDNA è necessaria perché, rispetto al DNA, l'mRNA è molto instabile e durante gli esperimenti sarebbe rapidamente degradato dall'azione degli ioni sodio, delle nucleasi cellulari o di quelle liberate dalla nostra pelle (in genere quando si manipola mRNA si usano i guanti impermeabili ai liquidi).
La scelta della genoteca da utilizzare dipende dalle informazioni che vogliamo ottenere. Se interessa conoscere la sequenza nucleotidica della parte codificante del gene al fine di dedurre da essa quella aminoacidica della

proteina, è più semplice utilizzare una genoteca di cDNA (è composta da molti meno cloni rispetto a quella genomica) ed anche perché dal cDNA avremo solo la sequenza codificante, senza introni. La presenza degli introni può rendere più laboriosa la definizione della sequenza della proteina. Tuttavia la genoteca di cDNA è limitata ai geni espressi nel tessuto dal quale è stato estratto l'mRNA totale. Se interessa conoscere la struttura del gene (promotore, stop di trascrizione, introni ed esoni) è necessario costruire una genoteca genomica. Una genoteca genomica include più del 90% di DNA di quella di cDNA pertanto richiede più lavoro per costruirla e vagliarla.
Per costruire una genoteca genomica occorre avere frammenti di DNA di dimensioni sufficientemente piccole da poterli poi inserire in vettori capaci di contenere frammenti di DNAds. I vettori (plasmidi, fagi, cosmidi e cromosomi) sono DNAds di varia origine, capaci di accettare frammenti di DNAds esogeno di dimensioni diverse (vedere tabella qui sotto) e di replicarlo in opportune cellule competenti che in relazione al tipo di vettore possono essere batteriche od eucariotiche (lievito).

Vettori	numero di coppie di basi trasportabili dal vettore
plasmide	10^4
batteriofago	$1,5x10^4$
cosmide	$4,5x10^4$
BAC	$1,5x10^5$
YAC	$2x10^6$

I plasmidi sono piccole porzioni circolari (minicromosomi) di DNAds extracromosomico presenti normalmente nei batteri. I fagi sono virus batterici ed i cosmidi sono plasmidi costruiti di piccole dimensioni in modo da poter accettare inserti di DNA più lunghi dei normali plasmidi. I cosmidi hanno l'origine di replicazione ed il gene di resistenza ad un antibiotico preso dai plasmidi e le estremità coesive (da ciò il loro nome) del fago lambda che permettono di impacchettare il costrutto (cosmide e inserto) nella capsula virale. BAC (Bacterial Artificial Chromosome) ed YAC (Yeast Artificial Chromosome) sono cromosomi modificati con caratteristiche molecolari che ne permettono la replicazione nelle cellule in cui vengono transfettate, rispettivamente batteriche e di lievito. I BAC sono plasmidi di grosse dimensioni e con altre caratteristiche strutturali che li rendono simili ai cromosomi batterici, gli YAC hanno il DNA del centromero e dei telomeri ed una sequenza autonoma di replicazione (ARS) del lievito ed essendo cromosomi eucariotici permettono l'inserimento di frammenti di DNAds di grosse dimensioni. Per costruire genoteche di genomi di grandi dimensioni, come quello umano, sono stati usati i cromosomi YAC e BAC.

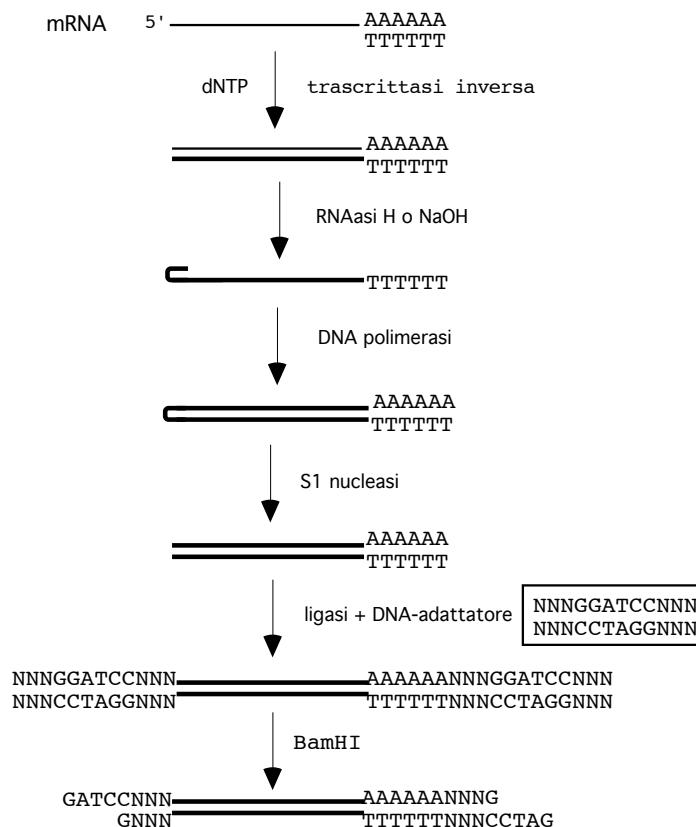



Figura 1-11. Sintesi del cDNA e legatura di adattatori per operare la sua clonazione.
La figura mostra schematicamente i passaggi necessari per sintetizzare un cDNA: conversione in DNA complementare (cDNA) e aggiunta di "adattatori" (linker) che sono frammenti di DNA con estremi netti contenenti la sequenza di un sito di restrizione (in figura BamHI) al fine di poter clonare il cDNA. L'mRNA viene estratto dalle cellule in coltura o dai tessuti, viene purificato mediante cromatografia di affinità utilizzando una resina fatta di un supporto insolubile di cellulosa alla quale sono stati legati covalentemente oligonucleotidi di deossitimidina (oligo-dT). L'estratto cellulare di RNA totale (rRNA, tRNA ed mRNA) viene fatto passare sulla colonna di cellulosa-oligo-dT. Gli mRNA che contengono il poliA (nell'uomo tutti gli mRNA esclusi quelli delle varianti di replicazione degli istoni) sono trattenuti sulla colonna perché i poliA si ibridano agli oligo-dT, mentre gli altri RNA passano attraverso la resina. Successivamente con opportune soluzioni (es. bassa concentrazione di sali) gli mRNA sono dissociati dalla colonna (eluiti) e quindi raccolti. L'eluato conterrà tutti gli mRNA presenti nelle cellule nello stato fisiologico in cui si trovavano quando sono state lisate (distrutte) per estrarre gli mRNA. La trascrittasi inversa viene incubata con gli mRNA, un oligo-dT (primer) ed i quattro desossinucleotiditrifosfati (dNTP). L'enzima catalizza la sintesi del primo filamento di cDNA ad ogni mRNA presente nell'incubazione formando un ibrido DNA-RNA. Successivamente l'mRNA è degradato con RNAasi H o NaOH (ambedue non alterano il DNA, l'RNAasi H degrada specificamente i filamenti di RNA ibridati a filamenti di DNA). Il filamento di cDNA tende a ripiegarsi indietro formando una breve struttura a forcella che è utilizzata dall'enzima DNA-polimerasi per sintetizzare il secondo filamento di cDNA. La breve ansa al 5' è tagliata con l'enzima S1 nucleasi, enzima che taglia specificamente i filamenti e regioni di DNA non accoppiate. Agli estremi del cDNA neosintetizzato vengono legati covalentemente due adattatori che contengono un sito di restrizione (in figura BamHI) e successivamente il cDNA viene digerito con lo stesso enzima di restrizione al fine di avere gli estremi coesivi e quindi per poterlo ricombinare con il DNA del vettore digerito con lo stesso enzima. Al termine della procedura in una unica provetta si hanno i cDNA originati da tutti gli mRNA presenti in un tipo di cellula/tessuto, essi possono essere inseriti nelle copie di uno stesso vettore (es. plasmide figura 1-7b) ed i vari cloni costituiscono una genoteca di cDNA di quel dato tipo di cellula/tessuto (figura 1-13). (ridisegnato e modificato da figura 7-1 di Recombinant DNA, Watson J.D., Gilman M., Witkowski J., Zoller M. (1992) Scientific American Books, 2nd ed., e figura 4.9 di Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK).

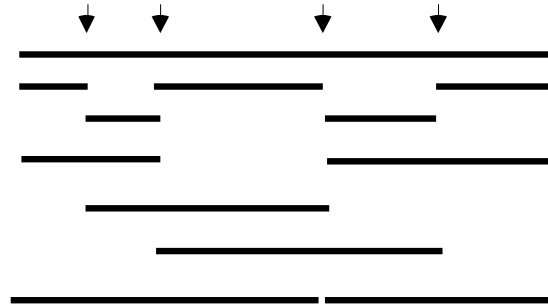



Figura 1-12. Digestione incompleta di un frammento di DNA con un enzima endonucleasi di restrizione. Le frecce puntano i siti di restrizione presenti nel frammento di DNA. Sono indicati alcuni frammenti del DNA che possono formarsi per digestione incompleta del frammento di DNA. Le prime due linee di frammenti mostrano i frammenti che sarebbero prodotti se la digestione fosse completa.

I cDNA essendo originati da mRNA hanno dimensioni relativamente piccole e possono essere inseriti come tali in plasmidi dopo averli muniti di adattatori (figura. 1-11). Il DNA genomico, che è di notevoli dimensioni (nell'uomo anche i cromosomi più piccoli sono costituiti da decine di milioni di nucleotidi), viene digerito (frammentato) mediante enzimi endonucleasi di restrizione. Poiché ogni sito di restrizione è ampiamente distribuito nel genoma, l'azione dell'enzima (digestione) per tempi ben definiti determina la scissione del DNA in frammenti delle giuste dimensioni (figura 1-12).
I frammenti di DNA genomico o di cDNA sono mescolati con il DNA di un vettore, avente estremità di DNAss complementari a quelle del DNA di interesse perché trattato con lo stesso enzima di restrizione e quindi legati covalentemente mediante una reazione catalizzata dall'enzima ligasi (figure 1-4, 1-7 e 1-13) ed il prodotto di sintesi è detto costrutto. I vari costrutti sono inseriti in cellule competenti (batteriche o cellule eucariotiche), capaci di ricevere il DNA del vettore e di farlo replicare (figure 1-7b e 1-13).
L'inserimento a caso di differenti frammenti di DNA nelle diverse copie di un plasmide è detto "shotgun cloning" (clonazione come un colpo di fucile a pallini). Quando il vettore è un batteriofago, l'infezione dei batteri provoca la replicazione del costrutto ed anche la lisi delle cellule batteriche. Le colonie con cellule lisate sono chiamate placche (figura 1-13).

Tecnologie per il vaglio delle genoteche

<u>Vaglio della genoteca mediante una sonda radioattiva</u>.
Per individuare il clone contenente il costrutto (vettore-inserto) di interesse si può vagliare (screening) la genoteca utilizzando come sonda (probe) un frammento di DNA avente la stessa sequenza di almeno una parte del DNA (genomico o cDNA) che abbiamo interesse a clonare (figura 1-14).
La sonda può essere: a) un insieme di oligonucleotidi sintetici diversi in sequenza detti degenerati. Da piccole quantità di una proteina pura si può

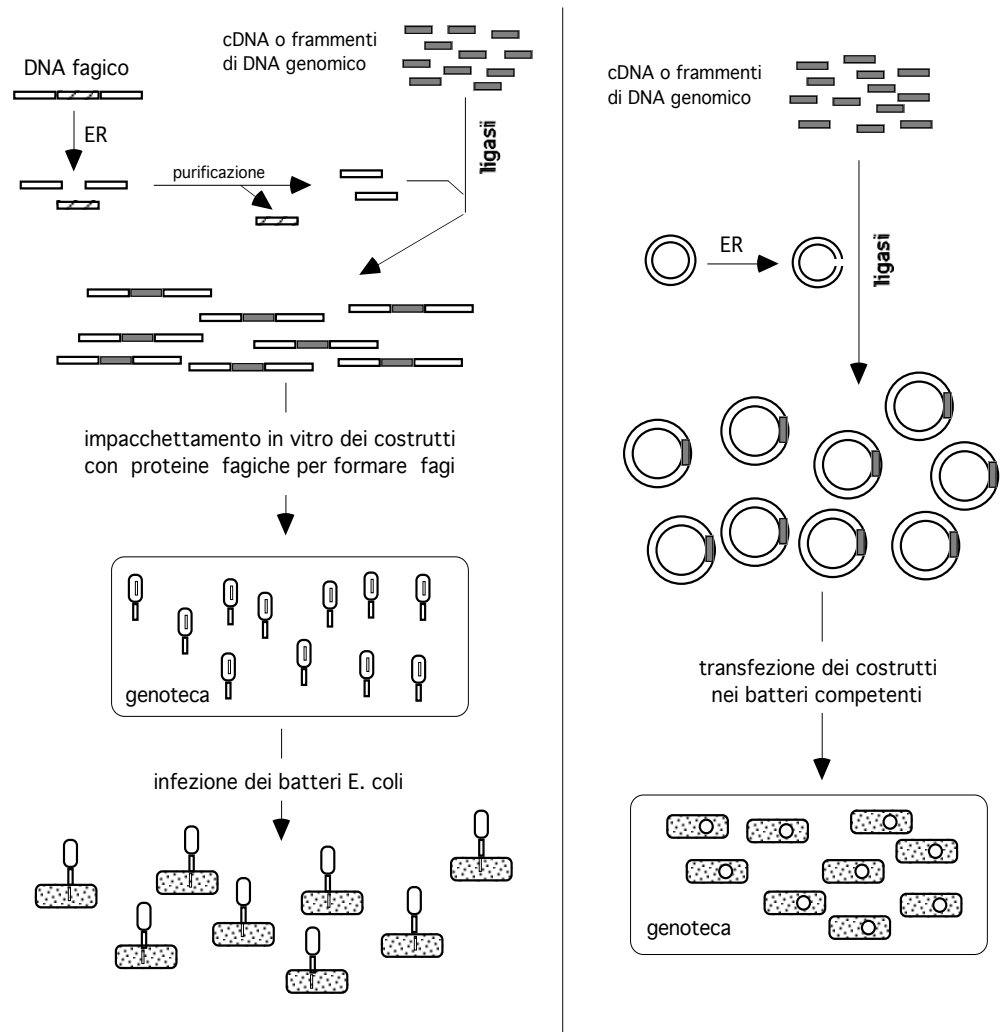



Figura 1-13. Schema della costruzione di una genoteca in batteriofagi (sinistra) ed una in plasmidi (destra).
La genoteca in batteriofagi è costituita da un insieme di fagi sintetici aventi singoli costrutti (DNAfagico-DNAinserto-DNAfagico). Per costruire questo costrutto il DNA del fago viene digerito con un enzima di restrizione (ER) al fine di eliminare la parte centrale non necessaria per la replicazione e purificare le estremità. Le estremità verranno poi legate al DNA dell'inserto (precedentemente digerito con lo stesso ER). La genoteca include un insieme di fagi ciascuno dei quali (presenti in più copie) contiene un costrutto diverso nell'inserto. Gli inserti sono di cDNA sintetizzati da tutti gli mRNA di un dato tessuto o frammenti di tutto il DNA genomico. Con i fagi verranno infettati i batteri competenti a ricevere il fago e a replicarne il DNA. La genoteca in plasmidi è costituita da una popolazione di batteri contenenti singoli costrutti (plasmide-inserto). La genoteca in plasmidi include un insieme batteri ciascuno dei quali (presenti in più copie) contiene un costrutto diverso nell'inserto.

---

ottenere la sequenza di un peptide di 10-20 aminoacidi dell'NH_2-terminale utilizzando analizzatori di sequenze aminoacidiche capaci di operare su poche pmoli di proteina, corrispondenti a poche decine di nanogrammi. Talvolta l'NH_2-terminale è bloccato (il primo aminoacido ha il suo gruppo amminico acetilato o metilato) e non può essere sottoposto ad analisi di sequenza del NH_2-

terminale. Occorre eliminare il gruppo che blocca o digerire parzialmente la proteina con un enzima proteolitico (es. tripsina), purificare i peptidi prodotti e determinare la loro sequenza mediante un sequenziatore automatico. Utilizzando il codice genetico, dalla sequenza aminoacidica del peptide(i) si deduce quella nucleotidica che lo codifica. Tuttavia, a causa della degenerazione del codice, si hanno più specie di sequenze nucleotidiche che codificano il peptide. Solo la metionina e triptofano sono codificati da una unica tripletta, mentre 9 aminoacidi sono codificati da due, 1 da tre, 5 da quattro e 3 da sei triplette diverse. E' quindi necessario sintetizzare: a) più oligonucleotidi codificanti il peptide con la speranza di indovinare quello con la sequenza nucleotidica identica o con alta similarità con quella naturale codificante il peptide stesso; b) un cDNA (o parte di esso) che viene utilizzato per vagliare una genoteca genomica al fine di clonare il gene cromosomico; c) un frammento di DNA (di gene o cDNA) di una isoforma (variante genetica) del gene di interesse della stessa specie (es. sonda e genoteca ambedue umane) o appartenente ad una specie diversa (sonda murina e genoteca umana), con la speranza che esista una sufficiente similarità di sequenza tra i due geni (quello noto e quello da ricercare). I batteri costituenti la genoteca, cioè l'insieme dei batteri contenenti il costrutto vettore-inserto, vengono sparsi sopra un terreno solido contenuto in una piastra (capsula di Petri: cilindro di plastica di diametro di circa 8cm e di altezza circa 1,5cm con coperchio della stessa forma cilindrica), in modo che singoli batteri possano formare colonie distanziate tra loro per evitare contaminazioni di cloni. Cioè che due o più batteri, contenenti lo stesso vettore ma diverso DNA-inserto, moltiplicandosi finiscano per mescolarsi tra loro dando l'apparenza di una singola colonia. Se la colonia non è omogenea (contiene più di un clone) non deve essere usata perché prelevando batteri da essa si hanno più specie di DNA-inserto. Lo scopo della clonazione è proprio quello di arrivare ad avere tanti batteri tutti originati da una singola cellula (clone) al fine di avere poi una preparazione più o meno grande di DNA (vettore-inserto) costituita solo di molecole identiche (omogenea). Il terreno su cui crescono i batteri è solido gelatinoso per evitare spostamenti di batteri/colonie durante le manipolazioni. Se la genoteca è costituita da batteriofagi, con i batteriofagi vengono infettati batteri competenti e le cellule infettate sono sparse sul terreno solido di varie piastre. Sulle piastre si formeranno placche (aree in cui il virus si è riprodotto ed ha lisato le cellule della colonia batterica) visibili ad occhio nudo. Parte di ciascuna colonia o placca è trasferita (blotting), per adesione da contatto, su un supporto solido (filtro di nitrocellulosa, nylon o di altro materiale plastico sintetico). Il filtro ha le stesse dimensioni del terreno di coltura, pertanto nel trasferimento le placche e le colonie mantengono le posizioni che avevano sul terreno. Le placche e colonie batteriche sul filtro sono trattate con soluzioni basiche chedistruggono le cellule e liberano il DNA dalle proteine. Il DNA è fissato sul filtro nella stessa posizione che avevano le placche e le colonie sul terreno. Il DNA viene poi sottoposto ad ibridazione contro la sonda radioattiva. Lo scopo del blotting, cioè del trasferimento delle colonie sul supporto solido è quello di

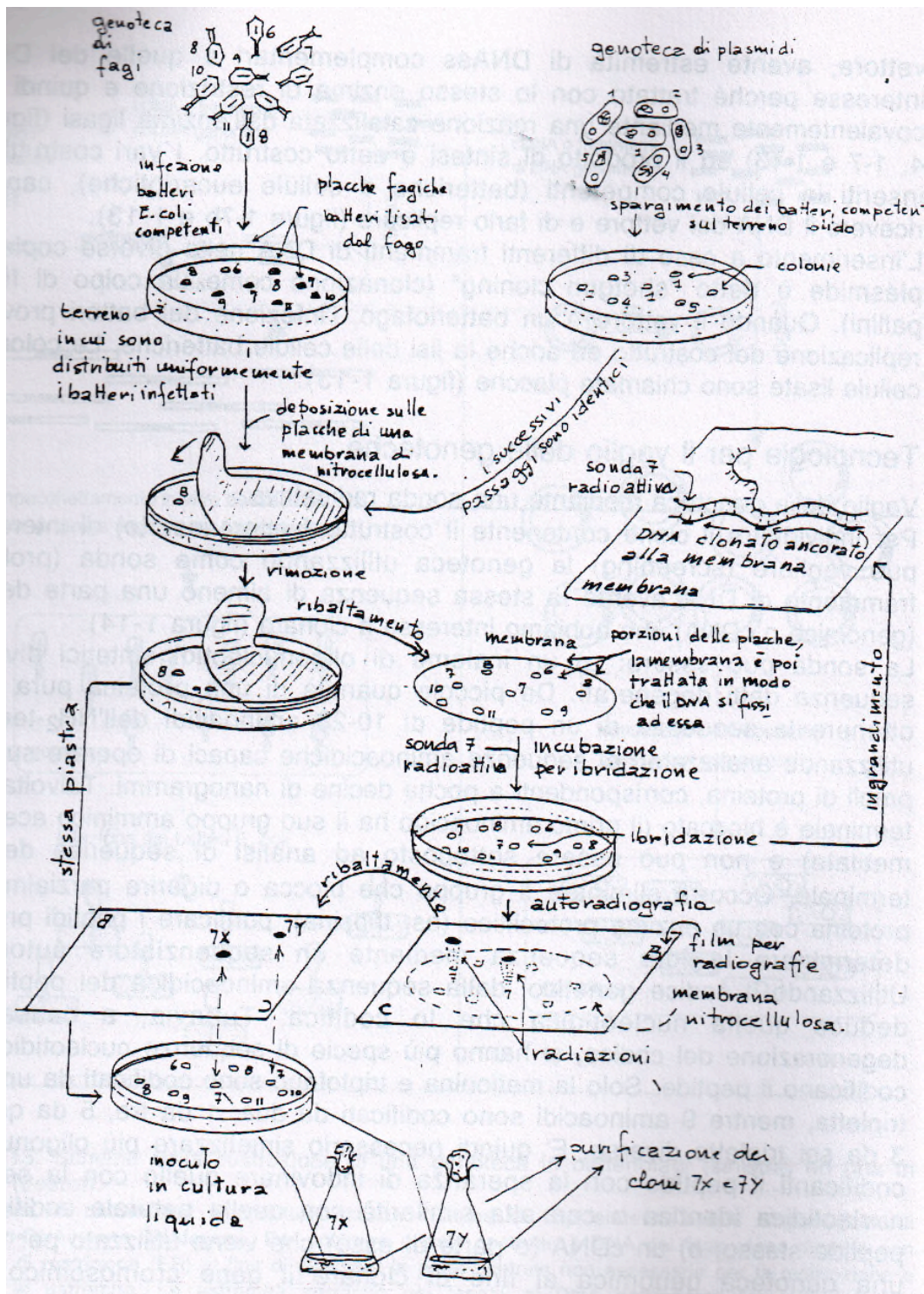



Figura 1-14. Vaglio di una genoteca in fagi e di una in plasmidi. Per il dettaglio vedere il testo (ridisegnato e modificato da Recombinant DNA, Watson J.D., Gilman M, Witkowski J., Zoller M. (1992) Scientific American Books, 2nd ed.).

poter operare la distruzione delle cellule, liberare il DNA dalle proteine ed eseguire l'ibridazione. Ciò non sarebbe possibile se le cellule fossero ancora nel terreno di coltura, fragile, gelatinoso e appiccicoso. Inoltre il blotting mantiene la disposizione che le colonie avevano nel terreno e questo è molto importante per le fasi successive del vaglio della genoteca. Per l'ibridazione, il filtro con il DNA fissato è immerso in un opportuno tampone ed, insieme al DNA della sonda, denaturato al calore di una T di 95°C. L'incubazione dura varie ore, tampone e temperatura di ibridazione sono scelti in modo che i singoli filamenti di DNA della sonda possano ibridarsi ai singoli filamenti del DNA dei cloni che hanno sequenze complementari. In questo modo con una sonda radioattiva e con una autoradiografia è possibile individuare le placche e le colonie batteriche che contengono il DNA del gene o cDNA di interesse. Le cellule batteriche trasferite sul filtro e sottoposte a lisi per liberare dalle proteine il loro DNA ed operare l'ibridazione, non sono più capaci di crescere, tuttavia la loro disposizione sul filtro è rivelata sul film autoradiografico. Si dispone il film autoradiografico sulla piastra dalla quale si era operato il trasferimento ed in questo modo si individua la posizione delle placche e delle colonie che sono risultate positive all'analisi Southern. I batteri delle colonie positive sono trasferiti in un terreno liquido dove crescono più rapidamente. I batteriofagi delle placche positive sono raccolti ed utilizzati per infettare colture liquide di batteri competenti. In questo modo (coltura in terreno liquido) si riesce ad ottenere molti batteri da cui estrarre una grande quantità di copie del costrutto (vettore-inserto) che è stato clonato. I costrutti saranno estratti dalle cellule e purificati al fine di analizzare ed utilizzare il DNA dell'inserto.

Quando la sequenza nucleotidica rivela che il frammento di cDNA clonato è incompleto (es. manca l'AUG e/o le triplette stop) si rende necessario vagliare di nuovo la genoteca. Per fare ciò risulta utile usare come sonda il frammento clonato nell'intento di individuare cloni sfuggiti alla sonda, precedentemente usata, perché mancante di un tratto di DNA presente nel frammento neoclonato.

Nella costruzione delle genoteche di cDNA, mRNA poco rappresentati nella cellula possono sfuggire alle reazioni di conversione in cDNA o dare poco prodotto-cDNA che è successivamente perso nelle manipolazioni per la costruzione della genoteca. Può anche accadere che i geni siano espressi in un particolare momento ancora ignoto (es. di notte per alcuni geni coinvolti nella spermatogenesi) e l'estrazione del mRNA sia fatta nelle normali ore di lavoro.

Vagliando una genoteca genomica in genere un singolo clone include solo un frammento del gene di interesse e per definire la sequenza di un gene occorre ripetere le operazioni (vaglio-sequenza) più volte. La procedura è detta "camminare sul cromosoma" (chromosome walking, figura 4-2). Per la clonazione di una genoteca genomica è importante che il digerito del DNA genomico includa frammenti diversi in numero di basi di DNA includenti una stessa regione di DNA. Ciò risulta da una digestione volutamente incompleta del DNA cromosomico ottenuta con uno o due enzimi di restrizione.

Poiché nella soluzione di digestione sono presenti più copie di DNA di uno stesso cromosoma, uno stesso sito di restrizione (stesso locus) casualmente verrà tagliato o saltato da uno stesso enzima di restrizione (figura 1-12). In questo modo, essendoci frammenti diversi con parti identiche delle loro sequenze, mediante sovrapposizione di queste sequenze è possibile ricostruire la sequenza completa del DNA della regione cromosomica (figura 3-12a e b).

Vaglio della genoteca mediante la tecnologia della PCR.
Quando possibile si preferisce vagliare la genoteca con la tecnologia della PCR analitica (figura 1-8) piuttosto che con una sonda (figura 1-14). Si raccolgono piccole quantità di batteri di ciascuna colonia (o placca) e si depositano in altrettante microprovette contenenti la miscela di incubazione per la PCR. La temperatura di circa 95°C distruggerà le strutture cellulari, libererà e denaturerà il DNA che poi, abbassando la T, si assocerà ai primer specifici per le regioni del DNA dell'inserto di interesse che si vuole clonare. Le incubazioni della PCR che daranno un singolo amplificato (avente la prevista dimensione visibile all'elettroforesi su gel di agarosio) indicheranno le colonie che contengono il costrutto con l'inserto di interesse.

Quando non è nota alcuna parte della sequenza aminoacidica della proteina o del DNA di interesse, e quindi non si ha la disponibilità di una sonda nucleotidica o di primer, si possono usare altre metodologie:

Vaglio delle genoteche di espressione

Le genoteche di espressione sono genoteche di cDNA il cui vettore oltre a replicare il DNA del costrutto produce mRNA e la proteina codificata dal cDNA utilizzando l'apparato sintetico della cellula ospite. Queste genoteche sono utilizzate per clonare cDNA ignoti quando non si conosce la sequenza del gene ne quella della proteina, pertanto non si possono utilizzare sonde nucleotidi o PCR analitica.
Tipicamente le genoteche di espressione sono di cDNA e non di geni perché con il cDNA è molto più semplice costruire la genoteca per sintetizzare la proteina. L'inserto è relativamente piccolo (molto più piccolo dell'intero gene) tuttavia include la sequenza per codificare l'intera proteina. Le genoteche genomiche, oltre ad avere un numero di cloni molto più grande di quello delle genoteche di cDNA, contengono inserti che sono parti di geni piuttosto che geni interi capaci di codificare l'intera proteina.

Questo spiega perché volendo clonare un gene del quale non si conosce la sequenza, né quella della proteina codificata, la via obbligata sia quella di clonare il cDNA e poi usare il cDNA come sonda per vagliare la genoteca genomica.

E' possibile transfettare un intero gene in cellule eucariotiche ed ottenere la sintesi della relativa proteina, tuttavia ciò è possibile perché il gene è già stato clonato e identifica la sua intera sequenza.

Vaglio della genoteca di espressione mediante anticorpi.
Quando si hanno anticorpi contro la proteina di interesse, si costruisce una genoteca di espressione che è costituita da vettori di espressione che sono capaci di replicare il proprio DNA e di trascrivere mRNA atti a codificare la sintesi delle proteine il cui cDNA sia stato inserito negli stessi vettori. Il cDNA deve essere inserito vicino ad un promotore con il corretto orientamento e con corretto quadro di lettura per la sintesi del mRNA che verrà tradotto nella proteina. La sintesi del mRNA e della proteina sono realizzate dai componenti molecolari (enzimi, ribosomi, fattori di regolazione, metaboliti, ecc.) delle cellule ospitanti i plasmidi. Il costrutto vettore-inserto è inserito in cellule competenti. Il vaglio della genoteca di espressione si effettua come quello fatto con una sonda nucleotidica (figura 1-14), con la differenza che la sonda è l'anticorpo che si lega specificamente alla proteina di interesse (analisi Western) ed il complesso proteina-anticorpo è poi rivelato dallo stesso anticorpo a cui è legata covalentemente una molecola fluorescente oppure con una reazione colorata prodotta da un substrato su cui agisce un enzima legato all'anticorpo o legato ad un altro anticorpo (secondario) che si lega al primo (primario) che interagisce con la proteina.
Si sono verificati casi in cui gli sperimentatori possedevano anticorpi contro una proteina ignota e di ignota attività molecolare e funzione fisiologica. Un esempio è dato dalla identificazione di proteine citotossiche, responsabili di gastriti croniche ed ulcere nell'uomo, prodotte dal batterio *Helycobacter pylori*. Si era osservata la presenza di particolari ceppi di questo batterio nello stomaco di pazienti portatori di gastrite o ulcera ed anche la presenza nel siero di questi stessi pazienti di anticorpi che reagivano con alcune proteine del *E. pylori.* Utilizzando questi anticorpi fu purificata una di queste proteine, poi utilizzata per produrre un anticorpo puro e specifico contro di essa. L'anticorpo così ottenuto è stato utilizzato per vagliare una genoteca di espressione ed in questo modo è stato possibile isolare il gene che codificava la proteina citotossica pur non conoscendo nulla della sua molecola o attività molecolare.

Vaglio della genoteca di espressione mediante il dosaggio dell'attività molecolare del prodotto genico.
Una genoteca di espressione può essere vagliata dosando l'attività molecolare della proteina di interesse. Se la proteina è un enzima si fa il dosaggio dell'attività catalitica, se è un recettore di un ormone si valuta la sua capacità di legare il relativo legante (ormone), se è un fattore di trascrizione la capacità di legarsi al promotore e attivare il gene.
La lisi delle cellule deve essere sufficientemente blanda da non denaturare la proteina, che perdendo la conformazione naturale perderebbe anche l'attività

molecolare. La denaturazione della proteina è meno drammatica per l'analisi Western perché gli anticorpi possono agire anche su proteine denaturate.

Vaglio della genoteca mediante complementazione funzionale.

Questo tipo di vaglio si basa sull'isolamento dei cloni capaci di correggere un difetto funzionale presente in un dato tipo di cellula.

Per questo vaglio occorre avere cellule portatrici del difetto funzionale e queste stesse cellule sono utilizzate come riceventi il costrutto (vettore-inserto) per la costruzione della genoteca di espressione. Questa particolare genoteca conterrà vettori che hanno inserito i cDNA sintezzati da tutti gli mRNA estratti dalle cellule normali dello stesso tessuto della stessa specie. Le cellule vengono poi cresciute in condizioni in cui solo le cellule aventi recuperato la funzione possano sopravvivere e crescere. Quindi queste cellule devono contenere il plasmide con inserito il cDNA codificante la proteina che ha restaurato la funzione persa. Il gene che codifica la proteina normale è detto della suscettibilità alla patologia perché se mutato diviene responsabile della patologia. In questo modo viene isolato il gene di interesse. Un esempio può essere un enzima che catalizza la sintesi di un composto necessario alla vita cellulare, e la mutazione del gene che codifica l'enzima causa una patologia. Le cellule che mancano di questo enzima devono essere cresciute in presenza del composto sopra indicato, altrimenti non crescono e muoiono. Queste stesse cellule sono transfettate con cDNA sintetizzato da mRNA estratti da cellule normali dello stesso tessuto e della stessa specie delle cellule patologiche. Se le cellule transfettate sono cresciute in assenza del composto necessario alla vita, tra tutte le cellule transfettate sopravvivranno solo le cellule mutate che avranno ricevuto il plasmide con inserito il cDNA che esprime l'enzima che catalizza la sintesi del composto necessario alla vita cellulare. Clonando queste cellule si clona e quindi purifica il cDNA che codifica la proteina naturale responsabile della suscettibilità alla patologia. Questo tipo di vaglio di genoteca è stato utilizzato per isolare un gene (dei forse quattro) responsabile dell'anemia di Fanconi (AN). AN è una mielopatia (della serie bianca del sangue), autosomica recessiva, caratterizzata da citopenia e sensibilità ad agenti che danneggiano il DNA, classificata (come tentativo) "patologia dei sistemi di riparazione del DNA". In conseguenza dell'instabilità dei cromosomi i portatori di questa patologia hanno la predisposizione a sviluppare tumori maligni. Per isolare il gene sono state utilizzate cellule prelevate da pazienti AN e mantenute in coltura. Queste cellule, se cresciute in presenza di mutageni chimici (mitomicina e diepossibutano), muoiono rapidamente. Le stesse cellule sono state usate come riceventi di una genoteca di espressione di cDNA ottenuti da mRNA di linfoblasti umani normali. Le cellule così transfettate sono state cresciute in presenza dei mutageni chimici sopra indicati, con il risultato di uccidere tutte le cellule escluse quelle che includevano il cDNA codificante la proteina capace di riparare il danno del DNA cellulare. In questo modo è stato clonato ed identificato il cDNA e quindi il gene di interesse conoscendo solo

l'alterazione funzionale (insufficiente resistenza ai mutageni) e senza avere alcuna informazioni sulla molecola responsabile della patologia.
La prima analisi a cui viene sottoposto il cDNA appena clonato è la determinazione della sequenza nucleotidica che fornisce importanti informazioni sulla struttura della proteina, se il DNA è genomico si ottengono anche importanti informazione sulla struttura della regione promotrice e quella in esoni ed introni del gene.

Primer per l'analisi della sequenza dei cloni delle genoteche di espressione.
Quando il vaglio delle genoteche di espressione è stato effettuato senza conoscere niente della sequenza del gene di interesse, il primer per iniziare a fare la sequenza del DNA di interesse è un primer universale sintetico. Esso ha una sequenza di circa 20b, complementare alla sequenza del DNA del vettore contiguo al 5' del DNA dell'inserto di interesse.
La clonazione di un cDNA di sequenza ignota è attuata traducendo *in vitro* la proteina da esso codificata, analizzando poi la proteina stessa mediante anticorpi, dosando la sua attività molecolare o con il saggio di complementazione funzionale (Capitolo 4).

Dalla sequenza nucleotidica del cDNA alla sequenza aminoacidica della proteina da esso codificata

La sequenza nucleotidica di un cDNA o quella degli esoni di un gene può essere tradotta nella corrispondente sequenza aminoacidica, utilizzando il codice genetico. La traduzione è fatta rapidamente mediante computer utilizzando specifici programmi oppure collegandosi via internet a siti web come http://www.expasy.org (ExPASy = Expert Protein Analysis System, Swiss Institute of Bioinformatics, CH).
Tutte le sequenze nucleotidiche conservate nelle banche dati sono scritte come sequenze di DNA, anche se si riferiscono a sequenze di mRNA o a genomi di virus ad RNA. Esse sono scritte indicando la sequenza 5'--->3' del solo filamento senso e per gli RNA le basi U sono sostituite con basi T. Pertanto il codice di inizio AUG appare come ATG, i codici di stop alla traduzione UAA, UAG e UGA appaiono come TAA, TAG e TGA e la sequenza segnale di poliadenilazione AUAAA come ATAAA.
Utilizzando il programma "Translate" del sito ExPASy è possibile scrivere la sequenza nucleotidica di un cDNA in una apposita finestra e, attivando il programma, si ottiene la traduzione della sequenza del DNAds nelle direzioni 5'-->3' dei due filamenti, anche se ne abbiamo scritto uno solo, il complementare è dedotto e tradotto automaticamente dal programma.
Viene letta la prima tripletta del frammento di DNA ed ordinatamente tutte le triplette a lei successive (figura 1-15). Questo modo ordinato di lettura della sequenza nucleotidica, tripletta per tripletta, è detto: quadro di lettura (in inglese: reading frame, rf). Poi automaticamente il programma considera il secondo ed il terzo quadro di lettura, iniziando a leggere rispettivamente dalla seconda e dalla terza base dalla prima tripletta. Lo stesso procedimento viene ripetuto leggendo il filamento complementare al primo nella direzione 5'-->3',

perché la sequenza nucleotidica di un cDNA non indica *per se* (cioè se non tradotta) il filamento senso. In genere dei sei quadri di lettura, cinque hanno nella loro sequenza ripetute triplette di stop alla traduzione (TAA, TAG, TGA), pertanto non possono codificare la proteina di interesse. Questi quadri sono chiamati quadri di lettura bloccati. Solo uno dei quadri di lettura ha una lunga sequenza non bloccata da stop. Questo quadro di lettura è detto quadro di lettura aperto (open reading frame, orf) e si assume che esso sia il quadro di lettura del mRNA che codifica la proteina di interesse, mentre ciò non è proponibile per gli altri cinque moduli bloccati (figura 1-15). Forse la presenza degli stop ripetuti è un meccanismo di difesa che la natura ha selezionato per evitare la sintesi casuale di proteine non volute.
Se la sequenza del quadro di lettura aperto corrisponde all'intero mRNA la sua identità sarà confermata dalla presenza della tripletta di inizio, ATG, che codifica la metionina, e, dopo una serie di triplette codificanti aminoacidi, una tripletta di stop. La traduzione della sequenza nucleotidica tra questi due codici, fornirà la sequenza aminoacidica completa della proteina. La sequenza del cDNA senso é ulteriormente confermata dalla presenza al 3' della sequenza consenso segnale di poliadenilazione (ATAAA) e dal poli-A stesso.
Il 1° codice di inizio della traduzione (ATG) al 5’ del quadro di lettura aperto (orf) si assume che codifichi la prima metionina della proteina codificata (vedere quadro 2 di figura 1-15). Il 1° ATG è incluso nella sequenza consenso di Kozak che favorisce l’inizio della traduzione da quel ATG al fine di evitare che la traduzione inizi su un altro ATG della stessa sequenza che sia o non sia nel corretto quadro di lettura. L’automatismo del programma elettronico traduce anche le sequenze che non sono tradotte *in vivo*: la sequenza tra il 5’ del codice di inizio e quella al 3’ del codice di stop (figura 1-15, quadro 2) e così le sequenze tra i segnali di stop dei quadri di lettura bloccati (figura 1-15, quadri 1, 2-6).
A queste regole esistono eccezioni. In alcuni mRNA i codici di inizio non sono ATG ma ACG, CTG, e GTG che codificano rispettivamente Thr, Leu, Val, aminoacidi che saranno i primi aminoacidi della proteina codificata. Gli mRNA delle proteine istoniche, varianti di replicazione (sintetizzate durante la sintesi del DNA), non hanno poli-A. In alcuni mRNA, la traduzione non inizia dal 1° ATG del quadro di lettura aperto ma dal secondo, che a differenza del primo è incluso nella sequenza consenso di Kozac. Risulta così che la proteina sintetizzata nelle cellule è più corta di quella dedotta dalla sequenza del quadro di lettura aperto.
La sequenza aminoacidica ottenuta per traduzione simulata è poi confrontata elettronicamente con la sequenza aminoacidica della proteina o dei suoi peptidi mediante i programmi BLAST o Align disponibili nei siti web: http://www.expasy.org, http://www.ebi.ac.uk (ebi = European Bioinformatic Institute, GB), http://www.ncbi.nlm.nih.gov (ncbi = National Center for Biotecnology Information, USA).

Questo confronto è necessario per verificare che sia stato clonato il cDNA codificante la proteina di interesse ed esso è considerato l'unica verifica valida perché confronta la struttura (sequenza) delle due molecole (figura1-16).

Quando la sequenza della proteina o dei suoi peptidi è ignota perché il cDNA è stato clonato mediante anticorpi o clonazione funzionale, si assume che la sequenza del quadro di lettura aperto sia quella codificata dal mRNA dal quale è stato sintetizzato il cDNA. Si cerca la conferma facendo analisi (rispettivamente la reazione con gli anticorpi e il dosaggio dell'attività molecolare) sulla proteina sintetizzata *in vitro* dal cDNA al fine di verificare che essa risponda come la proteina naturale utilizzata per clonare il cDNA. Tuttavia, si può sempre pensare che esistano due proteine simili nell'attività molecolare ed in parte della sequenza e che l'anticorpo ed il dosaggio dell'attività molecolare non riescano a distinguerle perché egualmente attivi su ambedue le proteine. Se si vuole escludere questi dubbi occorre purificare la proteina da cellule e fare la sequenza aminoacidica di almeno alcuni dei suoi peptidi.

Dedurre la sequenza aminoacidica di una proteina dalla sequenza nucleotidica del relativo gene è più laborioso rispetto a dedurre la sequenza della stessa proteina con il cDNA, per la presenza di introni (sequenze non codificanti) intercalate con gli esoni (sequenze codificanti). Tuttavia, abbiamo visto che volendo clonare un gene utilizzando dati della proteina codificata (vedere prima vaglio delle genoteche e al capitolo 4 la clonazione funzionale) viene clonato prima il cDNA e poi il gene
Con la clonazione posizionale (capitolo 4) viene clonato direttamente il gene e per dedurre la sequenza della proteina (ed anche per sintetizzarla *in vitro*) si preferisce clonare il corrispondente cDNA utilizzando un suo esone o parte di esso come sonda o una analisi PCR specifica per un esone per vagliare la genoteca.

Figura 1-15. Pagine seguenti. Traduzione elettronica dei sei quadri di lettura di un cDNA.
La sequenza nucleotidica è scritta con caratteri minuscoli per distinguerla dalla sequenza proteica. Le triplette di inizio della traduzione sono sottolineate ed i codici di stop sono indicati come trattini nella sequenza aminoacidica. Il quadro di lettura 2 è aperto (orf) e codifica una sequenza aminoacidica di 280 aminoacidi, il suo codice di inizio (AUG), la 1ª metionina (M), il codice di stop (TGA) alla traduzione della proteina e la sequenza segnale di poliadenilazione (ATAAA) sono in grassetto. Gli aminoacidi indicati in corsivo prima della 1ª metionina non sono tradotti *in vivo* perché la traduzione inizia dal 1° AUG del quadro di lettura aperto. Gli altri quadri di lettura (1, 3, 4-6) sono bloccati ripetutamente da codici di stop alla traduzione.

```
5'3' Quadro 1, bloccato
cctgattcagcaggaagcataacagacaccaaccactatgctgtcagcagttgcccgggg
 P  D  S  A  G  S  I  T  D  T  N  H  Y  A  V  S  S  C  P  G
ctaccagggctggtttcatccctgtgctaggctttctgtgaggatgagcagcaccgggat
 L  P  G  L  V  S  S  L  C  -  A  F  C  E  D  E  Q  H  R  D
agacaggaagggcgtcctggctaaccgggtagccgtggtcacggggtccaccagtgggat
 R  Q  E  G  R  P  G  -  P  G  S  R  G  H  G  V  H  Q  W  D
cggctttgccatcgcccgacgtctggcccgggacggggcccacgtggtcatcagcagccg
 R  L  C  H  R  P  T  S  G  P  G  R  G  P  R  G  H  Q  Q  P
gaagcagcagaacgtggaccgggccatggccaagctgcagggggagggggctgagtgtggc
 E  A  A  E  R  G  P  G  H  G  Q  A  A  G  G  G  A  E  C  G
gggcattgtgtgccacgtggggaaggctgaggaccgggagcagctggtggccaaggccct
 G  H  C  V  P  R  G  E  G  -  G  P  G  A  A  G  G  Q  G  P
ggagcactgtgggggcgtcgacttcctggtgtgcagcgcaggggtcaaccctctggtagg
 G  A  L  W  G  R  R  L  P  G  V  Q  R  R  G  Q  P  S  G  R
gagcactctggggaccagtgagcagatctgggacaagatcctaagtgtgaacgtgaagtc
 E  H  S  G  D  Q  -  A  D  L  G  Q  D  P  K  C  E  R  E  V
cccagccctgctgctgagccagttgctgccctacatggagaacaggagggtgctgtcat
 P  S  P  A  A  E  P  V  A  A  L  H  G  E  Q  E  G  C  C  H
cctggtctcttccattgcagcttataatccagtagtggcgctgggtgtctacaatgtcag
 P  G  L  F  H  C  S  L  -  S  S  S  G  A  G  C  L  Q  C  Q
caagacagcgctgctgggtctcactagaacactggcattggagctggcccccaaggacat
 Q  D  S  A  A  G  S  H  -  N  T  G  I  G  A  G  P  Q  G  H
ccgggtaaactgcgtggttccaggaattatcaaaactgacttcagcaaagtgtttcatgg
 P  G  K  L  R  G  S  R  N  Y  Q  N  -  L  Q  Q  S  V  S  W
gaatgagtctctctggaagaacttcaaggaacatcatcagctgcagaggattggggagtc
 E  -  V  S  L  E  E  L  Q  G  T  S  S  A  A  E  D  W  G  V
agaggactgtgcaggaatcgtgtccttcctgtgctctccagatgccagctacgtcaacgg
 R  G  L  C  R  N  R  V  L  P  V  L  S  R  C  Q  L  R  Q  R
ggagaacattgcggtggcaggctactccactcggctctgagaggagtgggggcggctgcg
 G  E  H  C  G  G  R  L  L  H  S  A  L  R  G  V  G  A  A  A
tagctgtggtcccaggcccaggagcctgagggggtgtctaggtgatcatttggatctgga
 -  L  W  S  Q  A  Q  E  P  E  G  V  S  R  -  S  F  G  S  G
ggcagagtctgccattctgccagactagcaatttgggggcttactcatgctaggcttgag
 G  R  V  C  H  S  A  R  L  A  I  W  G  L  T  H  A  R  L  E
gaagaagaaaaacgcttcggcattctccttaggacttatctgcttgtagatttggctgat
 E  E  E  K  R  F  G  I  L  L  R  T  Y  L  L  V  D  L  A  D
ccaattaacatgtggggttcttggtgtgggtctggggagctgaaggattttatggagctg
 P  I  N  M  W  G  S  W  C  G  S  G  E  L  K  D  F  M  E  L
gtgctttggaggaatcttaagggaaaggagtagaagctcaggcctttgaaggatttcagc
 V  L  W  R  N  L  K  G  K  E  -  K  L  R  P  L  K  D  F  S
tcctcctctctgtaatttgtgctttaagcatttttttcctaaaataaactcaaatttat
 S  S  S  L  -  F  V  L  -  A  F  F  F  L  K  -  T  Q  I  Y
cctcaa
 P  Q
```

5'3' Quadro 2 aperto

```
                                                    1°
cctgattcagcaggaagcataacagacaccaaccactatgctgtcagcagttgcccggggc
 L  I  Q  Q  E  A  -  Q  T  P  T  T  M  L  S  A  V  A  R  G
taccagggctggtttcatccctgtgctaggctttctgtgaggatgagcagcaccgggata
 Y  Q  G  W  F  H  P  C  A  R  L  S  V  R  M  S  S  T  G  I
gacaggaagggcgtcctggctaaccgggtagccgtggtcacggggtccaccagtgggatc
 D  R  K  G  V  L  A  N  R  V  A  V  V  T  G  S  T  S  G  I
ggctttgccatcgcccgacgtctggcccgggacggggcccacgtggtcatcagcagccgg
 G  F  A  I  A  R  R  L  A  R  D  G  A  H  V  V  I  S  S  R
aagcagcagaacgtggaccgggccatggccaagctgcagggggaggggctgagtgtggcg
 K  Q  Q  N  V  D  R  A  M  A  K  L  Q  G  E  G  L  S  V  A
ggcattgtgtgccacgtggggaaggctgaggaccgggagcagctggtggccaaggccctg
 G  I  V  C  H  V  G  K  A  E  D  R  E  Q  L  V  A  K  A  L
gagcactgtgggggcgtcgacttcctggtgtgcagcgcaggggtcaaccctctggtaggg
 E  H  C  G  G  V  D  F  L  V  C  S  A  G  V  N  P  L  V  G
agcactctggggaccagtgagcagatctgggacaagatcctaagtgtgaacgtgaagtcc
 S  T  L  G  T  S  E  Q  I  W  D  K  I  L  S  V  N  V  K  S
ccagccctgctgctgagccagttgctgccctacatggagaacaggaggggtgctgtcatc
 P  A  L  L  L  S  Q  L  L  P  Y  M  E  N  R  R  G  A  V  I
ctggtctcttccattgcagcttataatccagtagtggcgctgggtgtctacaatgtcagc
 L  V  S  S  I  A  A  Y  N  P  V  V  A  L  G  V  Y  N  V  S
aagacagcgctgctgggtctcactagaacactggcattggagctggcccccaaggacatc
 K  T  A  L  L  G  L  T  R  T  L  A  L  E  L  A  P  K  D  I
cgggtaaactgcgtggttccaggaattatcaaaactgacttcagcaaagtgtttcatggg
 R  V  N  C  V  V  P  G  I  I  K  T  D  F  S  K  V  F  H  G
aatgagtctctctggaagaacttcaaggaacatcatcagctgcagaggattggggagtca
 N  E  S  L  W  K  N  F  K  E  H  H  Q  L  Q  R  I  G  E  S
gaggactgtgcaggaatcgtgtccttcctgtgctctccagatgccagctacgtcaacggg
 E  D  C  A  G  I  V  S  F  L  C  S  P  D  A  S  Y  V  N  G
gagaacattgcggtggcaggctactccactcggctctgagaggagtgggggcggctgcgt
 E  N  I  A  V  A  G  Y  S  T  R  L  -  E  E  W  G  R  L  R
agctgtggtcccaggcccaggagcctgagggggtgtctaggtgatcatttggatctggag
 S  C  G  P  R  P  R  S  L  R  G  C  L  G  D  H  L  D  L  E
gcagagtctgccattctgccagactagcaatttgggggcttactcatgctaggcttgagg
 A  E  S  A  I  L  P  D  -  Q  F  G  G  L  L  M  L  G  L  R
aagaagaaaaacgcttcggcattctccttaggacttatctgcttgtagatttggctgatc
 K  K  K  N  A  S  A  F  S  L  G  L  I  C  L  -  I  W  L  I
caattaacatgtggggttcttggtgtgggtctggggagctgaaggattttatggagctgg
 Q  L  T  C  G  V  L  G  V  G  L  G  S  -  R  I  L  W  S  W
tgctttggaggaatcttaagggaaaggagtagaagctcaggcctttgaaggatttcagct
 C  F  G  G  I  L  R  E  R  S  R  S  S  G  L  -  R  I  S  A
cctcctctctgtaatttgtgctttaagcatttttttcctaaaataaactcaaatttatc
 P  P  L  C  N  L  C  F  K  H  F  F  S  -  N  K  L  K  F  I
ctcaa
 L
```

5'3' Quadro 3, bloccato

```
cctgattcagcaggaagcataacagacaccaaccactatgctgtcagcagttgcccggggct
   -  F  S  R  K  H  N  R  H  Q  P  L  C  C  Q  Q  L  P  G  A
accagggctggtttcatccctgtgctaggctttctgtgaggatgagcagcaccgggatag
 T  R  A  G  F  I  P  V  L  G  F  L  -  G  -  A  A  P  G  -
acaggaagggcgtcctggctaaccgggtagccgtggtcacggggtccaccagtgggatcg
 T  G  R  A  S  W  L  T  G  -  P  W  S  R  G  P  P  V  G  S
gctttgccatcgcccgacgtctggcccgggacggggcccacgtggtcatcagcagccgga
 A  L  P  S  P  D  V  W  P  G  T  G  P  T  W  S  S  A  A  G
agcagcagaacgtggaccgggccatggccaagctgcaggggagggctgagtgtggcgg
 S  S  R  T  W  T  G  P  W  P  S  C  R  G  R  G  -  V  W  R
gcattgtgtgccacgtggggaaggctgaggaccgggagcagctggtggccaaggccctgg
 A  L  C  A  T  W  G  R  L  R  T  G  S  S  W  W  P  R  P  W
agcactgtgggggcgtcgacttcctggtgtgcagcgcaggggtcaaccctctggtaggga
 S  T  V  G  A  S  T  S  W  C  A  A  Q  G  S  T  L  W  -  G
gcactctggggaccagtgagcagatctgggacaagatcctaagtgtgaacgtgaagtccc
 A  L  W  G  P  V  S  R  S  G  T  R  S  -  V  -  T  -  S  P
cagccctgctgctgagccagttgctgccctacatggagaacaggagggtgctgtcatcc
 Q  P  C  C  -  A  S  C  C  P  T  W  R  T  G  G  V  L  S  S
tggtctcttccattgcagcttataatccagtagtggcgctgggtgtctacaatgtcagca
 W  S  L  P  L  Q  L  I  I  Q  -  W  R  W  V  S  T  M  S  A
agacagcgctgctgggtctcactagaacactggcattggagctggcccccaaggacatcc
 R  Q  R  C  W  V  S  L  E  H  W  H  W  S  W  P  P  R  T  S
gggtaaactgcgtggttccaggaattatcaaaactgacttcagcaaagtgtttcatggga
 G  -  T  A  W  F  Q  E  L  S  K  L  T  S  A  K  C  F  M  G
atgagtctctctggaagaacttcaaggaacatcatcagctgcagaggattggggagtcag
 M  S  L  S  G  R  T  S  R  N  I  I  S  C  R  G  L  G  S  Q
aggactgtgcaggaatcgtgtccttcctgtgctctccagatgccagctacgtcaacgggg
 R  T  V  Q  E  S  C  P  S  C  A  L  Q  M  P  A  T  S  T  G
agaacattgcggtggcaggctactccactcggctctgagaggagtgggggcggctgcgta
 R  T  L  R  W  Q  A  T  P  L  G  S  E  R  S  G  G  G  C  V
gctgtggtcccaggcccaggagcctgagggggtgtctaggtgatcatttggatctggagg
 A  V  V  P  G  P  G  A  -  G  G  V  -  V  I  I  W  I  W  R
cagagtctgccattctgccagactagcaatttgggggcttactcatgctaggcttgagga
 Q  S  L  P  F  C  Q  T  S  N  L  G  A  Y  S  C  -  A  -  G
agaagaaaaacgcttcggcattctccttaggacttatctgcttgtagatttggctgatcc
 R  R  K  T  L  R  H  S  P  -  D  L  S  A  C  R  F  G  -  S
aattaacatgtggggttcttggtgtgggtctggggagctgaaggattttatggagctggt
 N  -  H  V  G  F  L  V  W  V  W  G  A  E  G  F  Y  G  A  G
gctttggaggaatcttaagggaaaggagtagaagctcaggcctttgaaggatttcagctc
 A  L  E  E  S  -  G  K  G  V  E  A  Q  A  F  E  G  F  Q  L
ctcctctctgtaatttgtgctttaagcatttttttcctaaaataaactcaaatttatcc
 L  L  S  V  I  C  A  L  S  I  F  F  P  K  I  N  S  N  L  S
tcaa
 S
```

3'5' Quadro 4, bloccato

```
ttgaggataaatttgagtttattttaggaaaaaaaatgcttaaagcacaaattacagaga
 L  R  I  N  L  S  L  F  -  E  K  K  C  L  K  H  K  L  Q  R
ggaggagctgaaatccttcaaaggcctgagcttctactcctttcccttaagattcctcca
 G  G  A  E  I  L  Q  R  P  E  L  L  L  L  S  L  K  I  P  P
aagcaccagctccataaaatccttcagctccccagacccacaccaagaaccccacatgtt
 K  H  Q  L  H  K  I  L  Q  L  P  R  P  T  P  R  T  P  H  V
aattggatcagccaaatctacaagcagataagtcctaaggagaatgccgaagcgttttc
 N  W  I  S  Q  I  Y  K  Q  I  S  P  K  E  N  A  E  A  F  F
ttcttcctcaagcctagcatgagtaagcccccaaattgctagtctggcagaatggcagac
 F  F  L  K  P  S  M  S  K  P  P  N  C  -  S  G  R  M  A  D
tctgcctccagatccaaatgatcacctagacacccctcaggctcctgggcctgggacca
 S  A  S  R  S  K  -  S  P  R  H  P  L  R  L  L  G  L  G  P
cagctacgcagccgcccccactcctctcagagccgagtggagtagcctgccaccgcaatg
 Q  L  R  S  R  P  H  S  S  Q  S  R  V  E  -  P  A  T  A  M
ttctccccgttgacgtagctggcatctggagagcacaggaaggacacgattcctgcacag
 F  S  P  L  T  -  L  A  S  G  E  H  R  K  D  T  I  P  A  Q
tcctctgactccccaatcctctgcagctgatgatgttccttgaagttcttccagagagac
 S  S  D  S  P  I  L  C  S  -  -  C  S  L  K  F  F  Q  R  D
tcattcccatgaaacactttgctgaagtcagttttgataattcctggaaccacgcagttt
 S  F  P  -  N  T  L  L  K  S  V  L  I  I  P  G  T  T  Q  F
acccggatgtccttgggggccagctccaatgccagtgttctagtgagacccagcagcgct
 T  R  M  S  L  G  A  S  S  N  A  S  V  L  V  R  P  S  S  A
gtcttgctgacattgtagacacccagcgccactactggattataagctgcaatggaagag
 V  L  L  T  L  -  T  P  S  A  T  T  G  L  -  A  A  M  E  E
accaggatgacagcacccctcctgttctccatgtagggcagcaactggctcagcagcagg
 T  R  M  T  A  P  L  L  F  S  M  -  G  S  N  W  L  S  S  R
gctggggacttcacgttcacacttaggatcttgtcccagatctgctcactggtccccaga
 A  G  D  F  T  F  T  L  R  I  L  S  Q  I  C  S  L  V  P  R
gtgctccctaccagagggttgacccctgcgctgcacaccaggaagtcgacgcccccacag
 V  L  P  T  R  G  L  T  P  A  L  H  T  R  K  S  T  P  P  Q
tgctccagggccttggccaccagctgctcccggtcctcagccttccccacgtggcacaca
 C  S  R  A  L  A  T  S  C  S  R  S  S  A  F  P  T  W  H  T
atgcccgccacactcagcccctcccctgcagcttggccatggcccggtccacgttctgc
 M  P  A  T  L  S  P  S  P  C  S  L  A  M  A  R  S  T  F  C
tgcttccggctgctgatgaccacgtgggccccgtcccgggccagacgtcgggcgatggca
 C  F  R  L  L  M  T  T  W  A  P  S  R  A  R  R  R  A  M  A
aagccgatcccactggtggaccccgtgaccacggctacccggttagccaggacgcccttc
 K  P  I  P  L  V  D  P  V  T  T  A  T  R  L  A  R  T  P  F
ctgtctatcccggtgctgctcatcctcacagaaagcctagcacagggatgaaaccagccc
 L  S  I  P  V  L  L  I  L  T  E  S  L  A  Q  G  -  N  Q  P
tggtagccccgggcaactgctgacagcatagtggttggtgtctgttatgcttcctgctga
 W  -  P  R  A  T  A  D  S  I  V  V  G  V  C  Y  A  S  C  -
atcagg
 I  R
```

3'5' Quadro 5, bloccato

```
ttgaggataaatttgagtttattttaggaaaaaaaatgcttaaagcacaaattacagagag
  -  G  -  I  -  V  Y  F  R  K  K  N  A  -  S  T  N  Y  R  E
gaggagctgaaatccttcaaaggcctgagcttctactcctttcccttaagattcctccaa
 E  E  L  K  S  F  K  G  L  S  F  Y  S  F  P  L  R  F  L  Q
agcaccagctccataaaatccttcagctccccagacccacaccaagaaccccacatgtta
 S  T  S  S  I  K  S  F  S  S  P  D  P  H  Q  E  P  H  M  L
attggatcagccaaatctacaagcagataagtcctaaggagaatgccgaagcgtttttct
 I  G  S  A  K  S  T  S  R  -  V  L  R  R  M  P  K  R  F  S
tcttcctcaagcctagcatgagtaagcccccaaattgctagtctggcagaatggcagact
 S  S  S  S  L  A  -  V  S  P  Q  I  A  S  L  A  E  W  Q  T
ctgcctccagatccaaatgatcacctagacacccccctcaggctcctgggcctgggaccac
 L  P  P  D  P  N  D  H  L  D  T  P  S  G  S  W  A  W  D  H
agctacgcagccgcccccactcctctcagagccgagtggagtagcctgccaccgcaatgt
 S  Y  A  A  A  P  T  P  L  R  A  E  W  S  S  L  P  P  Q  C
tctccccgttgacgtagctggcatctggagagcacaggaaggacacgattcctgcacagt
 S  P  R  -  R  S  W  H  L  E  S  T  G  R  T  R  F  L  H  S
cctctgactccccaatcctctgcagctgatgatgttccttgaagttcttccagagagact
 P  L  T  P  Q  S  S  A  A  D  D  V  P  -  S  S  S  R  E  T
cattcccatgaaacactttgctgaagtcagttttgataattcctggaaccacgcagttta
 H  S  H  E  T  L  C  -  S  Q  F  -  -  F  L  E  P  R  S  L
cccggatgtccttgggggccagctccaatgccagtgttctagtgagacccagcagcgctg
 P  G  C  P  W  G  P  A  P  M  P  V  F  -  -  D  P  A  A  L
tcttgctgacattgtagacacccagcgccactactggattataagctgcaatggaagaga
 S  C  -  H  C  R  H  P  A  P  L  L  D  Y  K  L  Q  W  K  R
ccaggatgacagcacccctcctgttctccatgtagggcagcaactggctcagcagcaggg
 P  G  -  Q  H  P  S  C  S  P  C  R  A  A  T  G  S  A  A  G
ctggggacttcacgttcacacttaggatcttgtcccagatctgctcactggtccccagag
 L  G  T  S  R  S  H  L  G  S  C  P  R  S  A  H  W  S  P  E
tgctccctaccagagggttgacccctgcgctgcacaccaggaagtcgacgcccccacagt
 C  S  L  P  E  G  -  P  L  R  C  T  P  G  S  R  R  P  H  S
gctccagggccttggccaccagctgctcccggtcctcagccttccccacgtggcacacaa
 A  P  G  P  W  P  P  A  A  P  G  P  Q  P  S  P  R  G  T  Q
tgcccgccacactcagcccctcccccctgcagcttggccatggcccggtccacgttctgct
 C  P  P  H  S  A  P  P  P  A  A  W  P  W  P  G  P  R  S  A
gcttccggctgctgatgaccacgtgggccccgtcccgggccagacgtcgggcgatggcaa
 A  S  G  C  -  -  P  R  G  P  R  P  G  P  D  V  G  R  W  Q
agccgatcccactggtggaccccgtgaccacggctacccggttagccaggacgcccttcc
 S  R  S  H  W  W  T  P  -  P  R  L  P  G  -  P  G  R  P  S
tgtctatcccggtgctgctcatcctcacagaaagcctagcacagggatgaaaccagccct
 C  L  S  R  C  C  S  S  S  Q  K  A  -  H  R  D  E  T  S  P
ggtagccccgggcaactgctgacagcatagtggttggtgtctgttatgcttcctgctgaa
 G  S  P  G  Q  L  L  T  A  -  W  L  V  S  V  M  L  P  A  E
tcagg
 S
```

3'5' Quadro 6, bloccato

```
ttgaggataaatttgagtttattttaggaaaaaaaatgcttaaagcacaaattacagagagg
   E  D  K  F  E  F  I  L  G  K  K  M  L  K  A  Q  I  T  E  R
aggagctgaaatccttcaaaggcctgagcttctactcctttcccttaagattcctccaaa
 R  S  -  N  P  S  K  A  -  A  S  T  P  F  P  -  D  S  S  K
gcaccagctccataaaatccttcagctccccagacccacaccaagaaccccacatgttaa
 A  P  A  P  -  N  P  S  A  P  Q  T  H  T  K  N  P  T  C  -
ttggatcagccaaatctacaagcagataagtcctaaggagaatgccgaagcgtttttctt
 L  D  Q  P  N  L  Q  A  D  K  S  -  G  E  C  R  S  V  F  L
cttcctcaagcctagcatgagtaagcccccaaattgctagtctggcagaatggcagactc
 L  P  Q  A  -  H  E  -  A  P  K  L  L  V  W  Q  N  G  R  L
tgcctccagatccaaatgatcacctagacacccccctcaggctcctgggcctgggaccaca
 C  L  Q  I  Q  M  I  T  -  T  P  P  Q  A  P  G  P  G  T  T
gctacgcagccgcccccactcctctcagagccgagtggagtagcctgccaccgcaatgtt
 A  T  Q  P  P  P  L  L  S  E  P  S  G  V  A  C  H  R  N  V
ctccccgttgacgtagctggcatctggagagcacaggaaggacacgattcctgcacagtc
 L  P  V  D  V  A  G  I  W  R  A  Q  E  G  H  D  S  C  T  V
ctctgactccccaatcctctgcagctgatgatgttccttgaagttcttccagagagactc
 L  -  L  P  N  P  L  Q  L  M  M  F  L  E  V  L  P  E  R  L
attcccatgaaacactttgctgaagtcagttttgataattcctggaaccacgcagtttac
 I  P  M  K  H  F  A  E  V  S  F  D  N  S  W  N  H  A  V  Y
ccggatgtccttgggggccagctccaatgccagtgttctagtgagacccagcagcgctgt
 P  D  V  L  G  G  Q  L  Q  C  Q  C  S  S  E  T  Q  Q  R  C
cttgctgacattgtagacacccagcgccactactggattataagctgcaatggaagagac
 L  A  D  I  V  D  T  Q  R  H  Y  W  I  I  S  C  N  G  R  D
caggatgacagcacccctcctgttctccatgtagggcagcaactggctcagcagcagggc
 Q  D  D  S  T  P  P  V  L  H  V  G  Q  Q  L  A  Q  Q  Q  G
tggggacttcacgttcacacttaggatcttgtcccagatctgctcactggtccccagagt
 W  G  L  H  V  H  T  -  D  L  V  P  D  L  L  T  G  P  Q  S
gctccctaccagagggttgacccctgcgctgcacaccaggaagtcgacgcccccacagtg
 A  P  Y  Q  R  V  D  P  C  A  A  H  Q  E  V  D  A  P  T  V
ctccagggccttggccaccagctgctcccggtcctcagccttccccacgtggcacacaat
 L  Q  G  L  G  H  Q  L  L  P  V  L  S  L  P  H  V  A  H  N
gcccgccacactcagcccctcccccctgcagcttggccatggcccggtccacgttctgctg
 A  R  H  T  Q  P  L  P  L  Q  L  G  H  G  P  V  H  V  L  L
cttccggctgctgatgaccacgtgggccccgtcccgggccagacgtcgggcgatggcaaa
 L  P  A  A  D  D  H  V  G  P  V  P  G  Q  T  S  G  D  G  K
gccgatcccactggtggaccccgtgaccacggctacccggttagccaggacgcccttcct
 A  D  P  T  G  G  P  R  D  H  G  Y  P  V  S  Q  D  A  L  P
gtctatcccggtgctgctcatcctcacagaaagcctagcacagggatgaaaccagccctg
 V  Y  P  G  A  A  H  P  H  R  K  P  S  T  G  M  K  P  A  L
gtagccccgggcaactgctgacagcatagtggttggtgtctgttatgcttcctgctgaat
 V  A  P  G  N  C  -  Q  H  S  G  W  C  L  L  C  F  L  L  N
cagg
 Q
```

```
CAGGAAGCATCTCAGACACCAACCACTATGCTGTCAGCAGTTGCCCGGGGCTACCAGGGC  480
                           M  L  S  A  V  A  R  G  Y  Q  G    11

TGGTTTCATCCCTGTGCTAGGCTTTCTGTGAGGATGAGCAGCACCGGGATAGACAGGAAG  540
W  F  H  P  C  A  R  L  S  V  R  M  S  S  T  G  I  D  R  K    31

GGCGTCCTGGCTAACCGGGTAGCCGTGGTCACGGGGTCCACCAGTGGGATCGGCTTTGCC  600
G  V  L  A  N  R  V  A  V  V  T  G  S  T  S  G  I  G  F  A    51

ATCGCCCGACGTCTGGCCCGGGACGGGGCCCACGTGGTCATCAGCAGCCGGAAGCAGCAG  660
I  A  R  R  L  A  R  D  G  A  H  V  V  I  S  S  R  K  Q  Q    71

AACGTGGACCGGGCCATGGCCAAGCTGCAGGGGGAGGGGCTGAGTGTGGCGGGCATTGTG  720
N  V  D  R  A  M  A  K  L  Q  G  E  G  L  S  V  A  G  I  V    91

TGCCACGTGGGGAAGGCTGAGGACCGGGAGCAGCTGGTGGCCAAGGCCCTGGAGCACTGT  780
C  H  V  G  K  A  E  D  R  E  Q  L  V  A  K  A  L  E  H  C   111

GGGGGCGTCGACTTCCTGGTGTGCAGCGCAGGGGTCAACCCTCTGGTAGGGAGCACTCTG  840
G  G  V  D  F  L  V  C  S  A  G  V  N  P  L  V  G  S  T  L   131

GGGACCAGTGAGCAGATCTGGGACAAGATCCTAAGTGTGAACGTGAAGTCCCCAGCCCTG  900
G  T  S  E  Q  I  W  D  K  I  L  S  V  N  V  K  S  P  A  L   151

CTGCTGAGCCAGTTGCTGCCCTACATGGAGAACAGGAGGGGTGCTGTCATCCTGGTCTCT  960
L  L  S  Q  L  L  P  Y  M  E  N  R  R  G  A  V  I  L  V  S   171

TCCATTGCAGCTTATAATCCAGTAGTGGCGCTGGGTGTCTACAATGTCAGCAAGACAGCG  1020
S  I  A  A  Y  N  P  V  V  A  L  G  V  Y  N  V  S  K  T  A   191
                                          •••••••
CTGCTGGGTCTCACTAGAACACTGGCATTGGAGCTGGCCCCCAAGGACATCCGGGTAAAC  1080
L  L  G  L  T  R  T  L  A  L  E  L  A  P  K  D  I  R  V  N   211

TGCGTGGTTCCAGGAATTATCAAAACTGACTTCAGCAAAGTGTTTCATGGGAATGAGTCT  1140
C  V  V  P  G  I  I  K  T  D  F  S  K  V  F  H  G  N  E  S   231
                                                 •••••••
CTCTGGAAGAACTTCAAGGAACATCATCAGCTGCAGAGGATTGGGGAGTCAGAGGACTGT  1200
L  W  K  N  F  K  E  H  H  Q  L  Q  R  I  G  E  S  E  D  C   251

GCAGGAATCGTGTCCTTCCTGTGCTCTCCAGATGCCAGCTACGTCAACGGGGAGAACATT  1260
A  G  I  V  S  F  L  C  S  P  D  A  S  Y  V  N  G  E  N  I   271

GCGGTGGCAGGCTACTCCACTCGGCTCTGAGAGGAGTGGGGGCGGCTGCGTAGCTGTGGT  1320
A  V  A  G  Y  S  T  R  L  *                                     280
```

Figura 1-16. La figura mostra la proteina tradotta elettronicamente dal quadro di lettura aperto del cDNA (quadro 2, figura 1-15) . Le sequenze sottolineate ed in grassetto corrispondono esattamente alle sequenze dei peptidi della proteina di interesse. Con un altro programma simile al BLAST, ma cercando nella banca dati delle sequenze segnale (capitolo 3), si è individuata la sequenza segnale della migrazione della proteina nel nucleo (sottolineata) che conferma il dato morfologico ottenuto con il microscopio a fluorescenza indicante la presenza della proteina nel nucleo. Inoltre si sono individuate due sequenze segnale per la glicosilazione della proteine (punteggiate), tuttavia la proteina non risulta glicosilata in natura (le sequenze segnale possono essere mascherate da residui di altri aminoacidi e quindi non attaccabili dagli enzimi della glicosilazione).

Ricerca dell’identità di un gene o di un cDNA neoclonato mediante analisi elettronica della sua sequenza nucleotidica e della sequenza aminoacidica della proteina da esso codificata

Il confronto della sequenza nucleotidica del gene o del cDNA e quello della sequenza aminoacidica della proteina con le sequenze conservate nelle banche dati forniscono utili informazioni per identificare il gene e la proteina da esso codificata. Le banche dati e i programmi di gestione per analizzare le sequenze nucleotidiche e proteiche si trovano in siti web gestiti da vari centri di ricerca. I principali siti web sono: http://www.ebi.ac.uk e http://www.ncbi.nlm.nih.gov , precedentemente menzionati. I programmi di gestione per vagliare le sequenze nucleotidiche e proteiche sono BLAST e FASTA.

Analisi elettronica della sequenza nucleotidica del DNA

La sequenza nucleotidica del DNA di interesse inserita in uno dei programmi di gestione sopra indicati viene allineata, sequenza per sequenza, con tutte le sequenze nucleotidiche conservate negli archivi delle banche dati e le risultanti coppie di sequenze appaiono sul monitor allineate per ordine decrescente di identità di basi, esistenti tra la sequenza di interesse e le sequenze nella banca dati della specie di interesse (o se voluto anche di altre specie). Il programma simula decine di migliaia di ibridazioni molecolari e se risulta che la sequenza del gene neoclonato è diversa da tutte quelle presenti nella banca dati consultata vuol dire che il gene è un nuovo gene (nuovo per la comunità scientifica, i geni sono vecchi almeno quanto la specie umana, milioni di anni). Se la sequenza nucleotidica del gene neoclonato risulta identica ad una conservata nella banca dati, vuol dire che il gene di interesse era già stato identificato: clonato e sequenziato. La sequenza nucleotidica del gene neoclonato può risultare diversa per una o più basi (poste anche in parti non codificanti del gene) da quella del gene conservato nell’archivio elettronico. In questo caso, il DNA neoclonato può essere un allele del gene conservato nell’archivio elettronico o più improbabilmente una sua copia originata per duplicazione recente. Per verificare quale delle due possibilità sia vera occorre stabilire il locus fisico del due geni (neoclonato e già archiviato): se i loci dei due geni coincidono essi sono alleli di uno stesso gene, mentre se i loci sono diversi i due geni sono geni duplicati.

Quando le sequenze nucleotidiche di due geni (esoni ed introni) o dei loro cDNA sono identiche anche le sequenze aminoacidiche delle proteine da essi codificate risulteranno identiche. Nel secondo caso, geni con sequenze nucleotidiche simili, può risultare che le basi diverse siano nelle sequenze non codificanti del gene (introni, sequenze non tradotte al 5’ ed al 3’ del mRNA) oppure che siano negli esoni e per la degenerazione del codice genetico, la

Figura 1-17. Pagine seguenti. Allineamento con il programma Align (28) delle sequenze nucleotidiche di cDNA umani.
a) stesso cDNA, sequenze nucleotidiche identiche. b) cDNA con sequenze nucleotidiche simili. Identità: % di nucleotidi identici nella stessa posizione di sequenza, nt = nucleotidi; i nt identici sono indicati con due punti (:), quelli diversi con uno spazio vuoto e le delezioni con un trattino (-). Il trattino è inserito automaticamente dal programma in modo che l'allineamento non si arresti alla prima delezione ma continui individuando l'allineamento più esteso tra le due sequenze (corrispondente al maggior valore di identità). Il cDNA Hep27 è lo stesso delle figure 1-15 e 1-16, il cDNA RTDH1 è dell'enzima retinol deidrogenasi.

a)

```
cDNA Hep27                                  843 nt contro.
cDNA Hep27                                  843 nt

100% identità

10        20        30        40        50        60
ATGCTGTCAGCAGTTGCCCGGGGCTACCAGGGCTGGTTTCATCCCTGTGCTAGGCTTTCT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
ATGCTGTCAGCAGTTGCCCGGGGCTACCAGGGCTGGTTTCATCCCTGTGCTAGGCTTTCT
10        20        30        40        50        60

70        80        90       100       110       120
GTGAGGATGAGCAGCACCGGGATAGACAGGAAGGGCGTCCTGGCTAACCGGGTAGCCGTG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GTGAGGATGAGCAGCACCGGGATAGACAGGAAGGGCGTCCTGGCTAACCGGGTAGCCGTG
70        80        90       100       110       120

130       140       150       160       170       180
GTCACGGGGTCCACCAGTGGGATCGGCTTTGCCATCGCCCGACGTCTGGCCCGGGACGGG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GTCACGGGGTCCACCAGTGGGATCGGCTTTGCCATCGCCCGACGTCTGGCCCGGGACGGG
130       140       150       160       170       180

190       200       210       220       230       240
GCCCACGTGGTCATCAGCAGCCGGAAGCAGCAGAACGTGGACCGGGCCATGGCCAAGCTG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GCCCACGTGGTCATCAGCAGCCGGAAGCAGCAGAACGTGGACCGGGCCATGGCCAAGCTG
190       200       210       220       230       240

250       260       270       280       290       300
CAGGGGGAGGGGCTGAGTGTGGCGGGCATTGTGTGCCACGTGGGGAAGGCTGAGGACCGG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAGGGGGAGGGGCTGAGTGTGGCGGGCATTGTGTGCCACGTGGGGAAGGCTGAGGACCGG
250       260       270       280       290       300

310       320       330       340       350       360
GAGCAGCTGGTGGCCAAGGCCCTGGAGCACTGTGGGGGCGTCGACTTCCTGGTGTGCAGC
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GAGCAGCTGGTGGCCAAGGCCCTGGAGCACTGTGGGGGCGTCGACTTCCTGGTGTGCAGC
310       320       330       340       350       360

370       380       390       400       410       420
GCAGGGGTCAACCCTCTGGTAGGGAGCACTCTGGGGACCAGTGAGCAGATCTGGGACAAG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GCAGGGGTCAACCCTCTGGTAGGGAGCACTCTGGGGACCAGTGAGCAGATCTGGGACAAG
370       380       390       400       410       420

430       440       450       460       470       480
ATCCTAAGTGTGAACGTGAAGTCCCCAGCCCTGCTGCTGAGCCAGTTGCTGCCCTACATG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
ATCCTAAGTGTGAACGTGAAGTCCCCAGCCCTGCTGCTGAGCCAGTTGCTGCCCTACATG
430       440       450       460       470       480
```

```
490       500       510       520       530       540

GAGAACAGGAGGGGTGCTGTCATCCTGGTCTCTTCCATTGCAGCTTATAATCCAGTAGTG
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GAGAACAGGAGGGGTGCTGTCATCCTGGTCTCTTCCATTGCAGCTTATAATCCAGTAGTG
490       500       510       520       530       540

550       560       570       580       590       600
GCGCTGGGTGTCTACAATGTCAGCAAGACAGCGCTGCTGGGTCTCACTAGAACACTGGCA
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GCGCTGGGTGTCTACAATGTCAGCAAGACAGCGCTGCTGGGTCTCACTAGAACACTGGCA
550       560       570       580       590       600

610       620       630       640       650       660
TTGGAGCTGGCCCCCAAGGACATCCGGGTAAACTGCGTGGTTCCAGGAATTATCAAAACT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
TTGGAGCTGGCCCCCAAGGACATCCGGGTAAACTGCGTGGTTCCAGGAATTATCAAAACT
610       620       630       640       650       660

670       680       690       700       710       720
GACTTCAGCAAAGTGTTTCATGGGAATGAGTCTCTCTGGAAGAACTTCAAGGAACATCAT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
GACTTCAGCAAAGTGTTTCATGGGAATGAGTCTCTCTGGAAGAACTTCAAGGAACATCAT
670       680       690       700       710       720

730       740       750       760       770       780
CAGCTGCAGAGGATTGGGGAGTCAGAGGACTGTGCAGGAATCGTGTCCTTCCTGTGCTCT
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CAGCTGCAGAGGATTGGGGAGTCAGAGGACTGTGCAGGAATCGTGTCCTTCCTGTGCTCT
730       740       750       760       770       780

790       800       810       820       830       840
CCAGATGCCAGCTACGTCAACGGGGAGAACATTGCGGTGGCAGGCTACTCCACTCGGCTC
::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
CCAGATGCCAGCTACGTCAACGGGGAGAACATTGCGGTGGCAGGCTACTCCACTCGGCTC
790       800       810       820       830       840

TGA
:::
TGA
```

---

b)

cDNA Hep27 843 nt contro

cDNA RTDH-1 837 nt

70.6% identità

```
10        20        30        40        50        60
ATGCTGTCAGCAGTTGCCCGGGGCTACCAGGGCTGGTTTCATCCCTGTGCTAGGCTTTCT
::::    :  ::     ::::::: : ::: ::   : :    : : ::: ::  :::
ATGC----ACAAGG----CGGGGCTGCTAGGCCTCTGTGC----CCGGGCTTGGAATTCG
10            20        30            40

70        80        90        100       110
GTGAGGATGAGCAGCACCGGGATAGACAGGAAGGGCGTC-CTGGCTAACCGGGTAGCCGT
::: :::::  :::: ::::::: :::  :  :::   : :: :: ::   ::: ::: :
GTGCGGATGGCCAGCTCCGGGAT-GACCCGCCGGGACCCGCTCGCAAATAAGGTGGCCCT
50        60        70         80        90        100

120       130       140       150       160       170
GGTCACGGGGTCCACCAGTGGGATCGGCTTTGCCATCGCCCGACGTCTGGCCCGGGACGG
::: ::::  ::::::   ::::::::::: ::::::::::: ::: :::::: ::::::
GGTAACGGCCTCCACCGACGGGATCGGCTTCGCCATCGCCCGGCGTTTGGCCCAGGACGG
110       120       130       140       150       160

180       190       200       210       220       230
GGCCCACGTGGTCATCAGCAGCCGGAAGCAGCAGAACGTGGACCGGGCCATGGCCAAGCT
:::::: :::::: :::::::::::::::::::::: ::::::: :::  :::::: :::
GGCCCATGTGGTCGTCAGCAGCCGGAAGCAGCAGAATGTGGACCAGGCGGTGGCCACGCT
170       180       190       200       210       220
```

```
240       250       260       270       280       290
GCAGGGGGAGGGGCTGAGTGTGGCGGGCATTGTGTGCCACGTGGGGAAGGCTGAGGACCG
:::::::::::::::::: ::: ::::::  :::::::: ::::::::::: ::::::::
GCAGGGGGAGGGGCTGAGCGTGACGGGCACCGTGTGCCATGTGGGGAAGGCGGAGGACCG
230       240       250       260       270       280

300       310       320       330       340       350
GGAGCAGCTGGTGGCCAAGGCCCTGGAGCACTGTGGGGGCGTCGACTTCCTGGTGTGCAG
::::: ::::::::::: :::  :: :::    ::: ::  ::::  :::: :: : ::
GGAGCGGCTGGTGGCCATGGCTGTGAAGCTTCATGGAGGTATCGATATCCTAGTCTCCAA
290       300       310       320       330       340

360       370       380       390       400       410
CGCAGGGGTCAACCCTCTGGTAGGGAGCACTCTGGGGACCAGTGAGCAGATCTGGGACAA
:: :  ::::::::: :  : :: :::    :::    :: :::: :: : ::::::::
TGCTGCTGTCAACCCTTTCTTTGGAAGCCTAATGGATGTCACTGAGGAGGTGTGGGACAA
350       360       370       380       390       400

420       430       440       450        460       470
GATCCTAAGTGTGAACGTGAAGTCCCCAGCCCTGCTG-CTGAGCCAGTTGCTGCCCTACA
::  ::      : :: :::::: ::::::::::: :: :  :: :::: : ::::  : :
GACTCTGGACATTAATGTGAAGGCCCCAGCCCTGATGACAAAGGCAGT-GGTGCCAGAAA
410       420       430       440       450        460

480        490        500       510       520       530
TGGAGAAC--AGGAGG-GGTGCTGTCATCCTGGTCTCTTCCATTGCAGCTTATAATCCAG
:::::::   :::::: ::  : ::  :   : :: :::::::: ::::: :  : ::::
TGGAGAAACGAGGAGGCGGCTCAGTGGTGATCGTGTCTTCCATAGCAGCCTTCAGTCCAT
470       480       490       500       510       520

540       550       560       570       580       590
-TAGTGGCGCTGGGTGTCTACAATGTCAGCAAGACAGCGCTGCTGGGTCTCACTAGAACA
:  ::::  :  ::   ::::::::::: :: :::::  ::::::: :: :: :  ::
CTCCTGGCT-TCAGTCCTTACAATGTCAGTAAAACAGCCTTGCTGGGCCTGACCAAGACC
530        540       550       560       570       580

600       610       620       630       640       650
CTGGCATTGGAGCTGGCCCCCAAGGACATCCGGGTAAACTGCGTGGTTCCAGGAATTATC
:::::  : ::::::::::: : : ::::  :::: :::::: : :  :: ::: :::::
CTGGCCATAGAGCTGGCCCCAAGGAACATTAGGGTGAACTGCCTAGCACCTGGACTTATC
590       600       610       620       630       640

660       670        680       690       700       710
AAAACTGACTTCAGCAAAGTGTTTC--ATGGGAATGAGTCTCTCTGGAAGAACTTCAAGG
:: :::  ::::::::   :: :    ::::  : : :       :::: : : : :: :
AAGACTAGCTTCAGCAGGATGCTCTGGATGGACAAG-GAAAAAGAGGAA-AGCATGAAAG
650       660       670       680        690       700

720       730       740       750       760       770
AACATCATCAGCTGCAGAGGATTGGGGAGTCAGAGGACTGTGCAGGAATCGTGTCCTTCC
::    :  : : :    ::: : :: ::: ::::::: ::::: :: :::::::: ::::
AAACCCTGCGGATAAGAAGGTTAGGCGAGCCAGAGGATTGTGCTGGCATCGTGTCTTTCC
710       720       730       740       750       760

780       790       800       810       820
TGTGCTCTCCAGATGCCAGCTACGTCAACGGGGAGAACATTGC-GGTGGC---AGGCTAC
::::::::  :::::::::::: :::  ::::: :::: ::  :::::     :::   :
TGTGCTCTGAAGATGCCAGCTACATCACTGGGGA-AACAGTGGTGGTGGGTGGAGGAACC
770       780       790        800       810       820

830       840
TCCACTCGGCTCTGA
:  : :: ::::::
CCGTCCCGCCTCTGA
830
```

sequenza aminoacidica dedotta dalla sequenza nucleotidica risulti identica a quella codificata dal gene già conservato nelle banche dati. Nei due geni sono diverse solo le basi che non portano alla codifica di un diverso aminoacido (per cui si ha la stessa sequenza aminoacidica) oppure si ha qualche cambiamento di codice per cui si hanno proteine diverse per pochi aminoacidi. Queste sostituzioni possono essere conservative (gli aminoacidi hanno residui simili) per cui le due proteine molto probabilmente hanno la stessa attività molecolare. Queste sono in genere le differenze di sequenza osservate per gli alleli di uno stesso gene. Nei geni ripetuti, le differenze di sequenza nucleotidica e proteica sono maggiori fino a rendere il gene o la proteina inattiva come si osserva per gli pseudogeni.
Il grado di similarità dei geni è misurato come percentuale delle basi identiche nella stessa posizione della sequenza. Geni simili in sequenza sono detti omologhi se sono originati da uno stesso gene capostipite ed essi costituiscono una famiglia di geni, che include geni di specie diverse e/o geni diversi nel genoma di una stessa specie. L'omologia è una proprietà qualitativa e non è quindi misurabile (non esistono geni più o meno omologhi). Geni omologhi appartenenti a specie diverse sono detti ortologhi (es. i geni delle emoglobine dei vertebrati), la loro diversità di sequenza si è formata durante l'evoluzione. Geni omologhi presenti nel genoma di una stessa specie (es. geni dell'emoglobina umana: embrionale, fetale, adulta) sono detti paraloghi se la loro formazione è avvenuta per duplicazione di uno di essi dopo la formazione della specie. Geni omologhi codificano proteine omologhe aventi funzioni e strutture terziarie in genere molto simili (vedere dopo).

Analisi elettronica delle sequenze aminoacidiche delle proteine

Le sequenze delle proteine sono confrontate mediante allineamento con i programmi di gestione FASTA e BLAST, modificati per operare con simboli degli aminoacidi al posto di quelli delle basi del DNA.
L'analisi delle sequenze aminoacidiche risulta più semplice ed informativa della stessa analisi fatta sulla sequenza nucleotidica dei geni e cDNA che le codificano. Le cause della maggiore difficoltà a confrontare i geni e cDNA includono: la complessità della struttura dei geni in esoni ed introni (sequenze con funzioni diverse), la maggiore lunghezza delle sequenze nucleotidiche codificanti rispetto a quelle aminoacidiche (3 basi per ogni aminoacido). Inoltre le proteine sono le molecole che mediante la loro attività molecolare svolgono la funzione codificata dai geni, pertanto le modifiche operate sulla sequenza delle proteine (sostituzione o delezione di aminoacidi) informano immediatamente sulle alterazioni che le mutazioni possono avere sull'attività molecolare e funzione cellulare della proteina e quindi del gene. Le mutazioni osservate sul gene non sempre alterano la sequenza aminoacidica e ciò si verifica proprio analizzando la sequenza della proteina.
Le proteine aventi sequenze simili sono riunite in famiglie. Una famiglia di proteine è costituita da proteine simili in sequenza ed in genere con la stessa, o

con una molto simile, attività molecolare (es. le emoglobine e gli enzimi che catalizzano la stessa reazione). Gli enzimi di una stessa specie, appartenenti ad una stessa famiglia, in genere hanno la stessa attività catalitica su gruppi atomici identici di substrati diversi. Ad esempio gli enzimi umani deidrogenasi a NAD/NADP appartenenti alla famiglia “alcool deidrogenasi a catena corta” (detti a catena corta perché il loro polipeptide, catena di aminoacidi, è più corto di quello dell’alcool deidrogenasi del fegato) ossidano il gruppo alcolico (≡C-OH) a gruppo carbonilico (=C=O) appartenenti a molecole molto diverse.
La similarità di sequenza dei rispettivi membri varia molto da famiglia a famiglia di proteine: le emoglobine umane (alfa, beta, gamma e delta) hanno circa il 40% di aminoacidi identici e gli enzimi alcool deidrogenasi a catena corta il 15-20%. Le sostituzioni degli amminoacidi modificano la proteina dotandola di nuove proprietà molecolari in relazione alle cellule in cui sono contenute. Es. alfa e beta emoglobina nell’adulto ed emoglobina gamma nel feto. L’emoglobina dei globuli rossi fetali è più affine all’ossigeno per poterla sottrarre all’emoglobina materna. Isoforme di uno stesso enzima si trovano in distretti subcellulari diversi per la presenza di sequenze segnale e di aminoacidi rispettivamente diversi che ne favoriscono la migliore solubilità nei due distretti.
In passato le proteine erano riunite in famiglie esclusivamente sulla base della identica o simile attività molecolare. Si conosceva la sequenza aminoacidica di poche proteine, data la difficoltà di purificarle e poi di determinarne la sequenza. L’avvento delle tecnologie del DNA ha permesso la clonazione dei cDNA codificanti molte proteine di funzione nota e quindi la conoscenza della sequenza aminoacidica delle stesse proteine. In questo modo veniva confermata la relazione tra “struttura e funzione” cioè che a funzioni simili corrispondevano strutture proteiche simili. Per struttura delle proteine si intende la struttura primaria che determina la struttura terziaria e, per funzione delle proteine si intende la loro attività molecolare (vedere appendice A). Tuttavia esistono eccezioni, enzimi che come sequenza appartengono ad una famiglia e come attività molecolare ad un’altra. Un esempio è dato dall’enzima UDP-galattosio epimerasi, che catalizza l’isomerizzazione del galattosio a glucosio. Esso è una epimerasi che come sequenza appartiene alla famiglia delle deidrogenasi a catena corta. La Natura è opportunista, se la casualità di una mutazione porta alla perdita dell’attività molecolare di un enzima e ne fa acquistare un’altra, e l’attività nuova è utile alla cellula, l’accetta e la mantiene anche se strutturalmente appartiene ad un'altra famiglia (la Natura non è razzista molecolare). La mutazione dell’enzima UDP-galattosio epimerasi è responsabile della patologia “galattosemia di tipo III”.
Proteine appartenenti a specie diverse aventi la stessa attività molecolare responsabile di una funzione indispensabile per la vita di ogni tipo di cellula (es. citocromo C) costituiscono famiglie includenti proteine aventi sequenze aminoacidiche simili e con alcuni residui aminoacidici, conservati nell'evoluzione, presenti in tutte le proteine e per questo detti "aminoacidi fossili". Si assume che questi aminoacidi siano conservati perché la loro sostituzione o perdita

causerebbe la perdita dell'attività molecolare senza alterarne la struttura terziaria (es. residui aminoacidici responsabili della catalisi degli enzimi) o perché causano la perdita della struttura terziaria. Tali aminoacidi fanno parte delle sequenze consenso di domini aventi attività molecolare (es. dominio di legame del cofattore, dominio contenente i gruppi catalitici, dominio di legame al legante o substrato) o di sequenze consenso responsabili della struttura terziaria della proteina. Pertanto nell'evoluzione viene mantenuta la funzione perché viene mantenuta la struttura della proteina. Il citocromo C è un enzima di 105 aminoacidi, mitocondriale della catena respiratoria molto conservato nell'evoluzione. Rispetto alla sequenza del citocromo dell'uomo, la scimmia Rhesus ha 1 solo aminoacido diverso, l'asino 11, la pecora, il maiale la vacca e la balena 10, il coniglio 9, il serpente a sonagli 14, il tonno 21, il baco da seta 31, il grano 43, il lievito di birra 45, Candida krusei (miceto) 51. Questa analisi indica che pur sostituendo circa il 50% degli aminoacidi della sequenza si ha ancora un citocromo C attivo. Tra gli aminoacidi conservati solo alcuni sono conservati in tutte le specie e questi sono definiti fossili, altri aminoacidi sono conservati ma non hanno una funzione altamente specifica e possono essere sostituiti con altri aventi un residuo simile (Glu <--> Asp, Val <--> Ile).Le famiglie delle proteine sono di due tipi: proteine appartenenti alla stessa specie o a specie diverse o dei due tipi insieme, es. le emoglobine dei mammiferi.

Similarità delle proteine

La similarità delle proteine è misurata come percentuale di identità di residui aminoacidici identici nella stessa posizione di sequenza (figura 1-18).

Quando si analizza la sequenza di una proteina, oltre alla valutazione percentuale dell'identità, si prendono in considerazione anche le sostituzioni conservative degli aminoacidi, cioè le sostituzioni in cui un aminoacido è sostituito con un altro avente un residuo simile per ingombro sterico e/o carica (es. Asp con Glu, stessa carica e molto simili stericamente). La piccola differenza molecolare in genere non altera l'attività molecolare della proteina, specialmente se il residuo aminoacidico è sulla superficie della proteina, in questo modo la mutazione è sfuggita alla selezione naturale. Geni omologhi (geni diversi discendenti da uno stesso gene capostipite) codificano proteine anch'esse definite omologhe. Il termine omologia è talvolta usato impropriamente al posto di similarità perché proteine omologhe sono anche simili. Ma non è necessariamente vero il contrario, che proteine con sequenze simili discendano da uno stesso capostipite, cioè che siano omologhe.

I geni non omologhi codificanti proteine simili, pur non discendendo da un comune gene progenitore, hanno avuto una evoluzione convergente verso una stessa funzione. La pressione evolutiva ha selezionato geni diversi verso la codificazione di proteine con attività molecolare simile od identica e, per avere una tale attività molecolare, le sequenze aminoacidiche, mutazione dopo mutazione, si sono modificate opportunamente e sono divenute simili al fine di potersi autoassemblare in strutture terziarie simili.

Figura 1-18. Pagine seguenti. Allineamento con il programma Align (28) delle sequenze amminoacidiche di proteine umane (pagine 85 e 86).
a) stessa proteina, sequenze amminoacidiche identiche.
b) nella pagina seguente: proteine con sequenze amminoacidiche simili. Identità = % di residui aminoacidici identici nella stessa posizione di sequenza, aa = aminoacidi, aa indicati sono indicati con due punti (:), quelli simili (sostituzioni conservative) con un punto (.) e le delezioni con un trattino (-).
La proteina Hep27 è la stessa delle figure 1-15 e 1-16. Per altri dettagli vedere la figura 1-17 ed il testo.

```
a)
Proteina Hep27                                     280 aa contro.
Proteina Hep27                                     280 aa

100% identità

               10        20        30        40        50        60
       MLSAVARGYQGWFHPCARLSVRMSSTGIDRKGVLANRVAVVTGSTSGIGFAIARRLARDG
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
       MLSAVARGYQGWFHPCARLSVRMSSTGIDRKGVLANRVAVVTGSTSGIGFAIARRLARDG
               10        20        30        40        50        60

               70        80        90       100       110       120
       AHVVISSRKQQNVDRAMAKLQGEGLSVAGIVCHVGKAEDREQLVAKALEHCGGVDFLVCS
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
       AHVVISSRKQQNVDRAMAKLQGEGLSVAGIVCHVGKAEDREQLVAKALEHCGGVDFLVCS
               70        80        90       100       110       120

              130       140       150       160       170       180
       AGVNPLVGSTLGTSEQIWDKILSVNVKSPALLLSQLLPYMENRRGAVILVSSIAAYNPVV
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
       AGVNPLVGSTLGTSEQIWDKILSVNVKSPALLLSQLLPYMENRRGAVILVSSIAAYNPVV
              130       140       150       160       170       180

              190       200       210       220       230       240
       ALGVYNVSKTALLGLTRTLALELAPKDIRVNCVVPGIIKTDFSKVFHGNESLWKNFKEHH
       ::::::::::::::::::::::::::::::::::::::::::::::::::::::::::::
       ALGVYNVSKTALLGLTRTLALELAPKDIRVNCVVPGIIKTDFSKVFHGNESLWKNFKEHH
              190       200       210       220       230       240

              250       260       270       280
       QLQRIGESEDCAGIVSFLCSPDASYVNGENIAVAGYSTRL
       ::::::::::::::::::::::::::::::::::::::::
       QLQRIGESEDCAGIVSFLCSPDASYVNGENIAVAGYSTRL
              250       260       270       280
```

```
b)
proteina Hep27                                        280 aa contro.
proteina RTDH                                         278 aa

59.2% identità
               10          20        30        40        50
       MLSAVARGYQGWFHPCARL--SVRMSSTGIDRKGVLANRVAVVTGSTSGIGFAIARRLAR
       : .:   :  :    :::   ::::.:.:. :.  :::.::.::.::.:::::::::::.
       MHKA---GLLGL---CARAWNSVRMASSGMTRRDPLANKVALVTASTDGIGFAIARRLAQ
                     10        20        30        40        50

       60        70        80        90       100       110
       DGAHVVISSRKQQNVDRAMAKLQGEGLSVAGIVCHVGKAEDREQLVAKALEHCGGVDFLV
       ::::::.:::::::::.:.: ::::::::.: :::::::::::.::: :..  ::.:.::
       DGAHVVVSSRKQQNVDQAVATLQGEGLSVTGTVCHVGKAEDRERLVAMAVKLHGGIDILV
           60        70        80        90       100       110

      120       130       140       150       160        170
       CSAGVNPLVGSTLGTSEQIWDKILSVNVKSPALLLSQLLPYMENRRG-AVILVSSIAAYN
        .:.:::. :: . ..:..::: :..:::.:::. . ..: ::.: : .:..::::::..
       SNAAVNPFFGSLMDVTEEVWDKTLDINVKAPALMTKAVVPEMEKRGGGSVVIVSSIAAFS
          120       130       140       150       160       170

       180       190       200       210       220       230
       PVVALGVYNVSKTALLGLTRTLALELAPKDIRVNCVVPGIIKTDFSKVFHGNESLWKNFK
       :  ... ::::::::::::.:::.::::..:::::..::.:::.::...  ..   ...:
       PSPGFSPYNVSKTALLGLTKTLAIELAPRNIRVNCLAPGLIKTSFSRMLWMDKEKEESMK
          180       190       200       210       220       230

       240       250       260       270        280
       EHHQLQRIGESEDCAGIVSFLCSPDASYVNGENIAVAGYS-TRL
       :  ...:.:: :::::::::::: ::::..::...:.: . .::
       ETLRIRRLGEPEDCAGIVSFLCSEDASYITGETVVVGGGTPSRL
          240       250       260       270
--------------------------------------------------------------------------------------
```

La costituzione delle banche dati, la facilità della loro consultazione ha semplificato lo studio/costituzione di famiglie di proteine che sono costituite proprio sulla base della identità/similitudine di sequenza aminoacidica delle proteine di una stessa specie e/o di specie diverse. Lo studio delle famiglie di proteine ha favorito anche lo studio dell'evoluzione molecolare.

L'analisi delle omologie di sequenza tra proteine effettuata consultando le banche dati ha dato anche risultati inaspettati. La comparazione della sequenza aminoacidica determinata facendo la sequenza di peptidi del fattore di crescita derivato delle piastrine (PDGF) con le sequenze aminoacidiche dedotte dalle sequenze nucleotidiche conservate in una banca dati ha permesso di scoprire una forte omologia del PDGF con quella di un oncogene virale (v-*sis*). In questo modo fu riscontrato per la prima volta che la sequenza di una proteina avente un ruolo nella normale replicazione cellulare era simile alla sequenza di una oncoproteina virale, induttrice di una eccessiva e patologica proliferazione cellulare. Se ne dedusse che gli oncogeni erano geni normali costituenti del genoma umano e che se mutati potevano causare cancerogenesi.

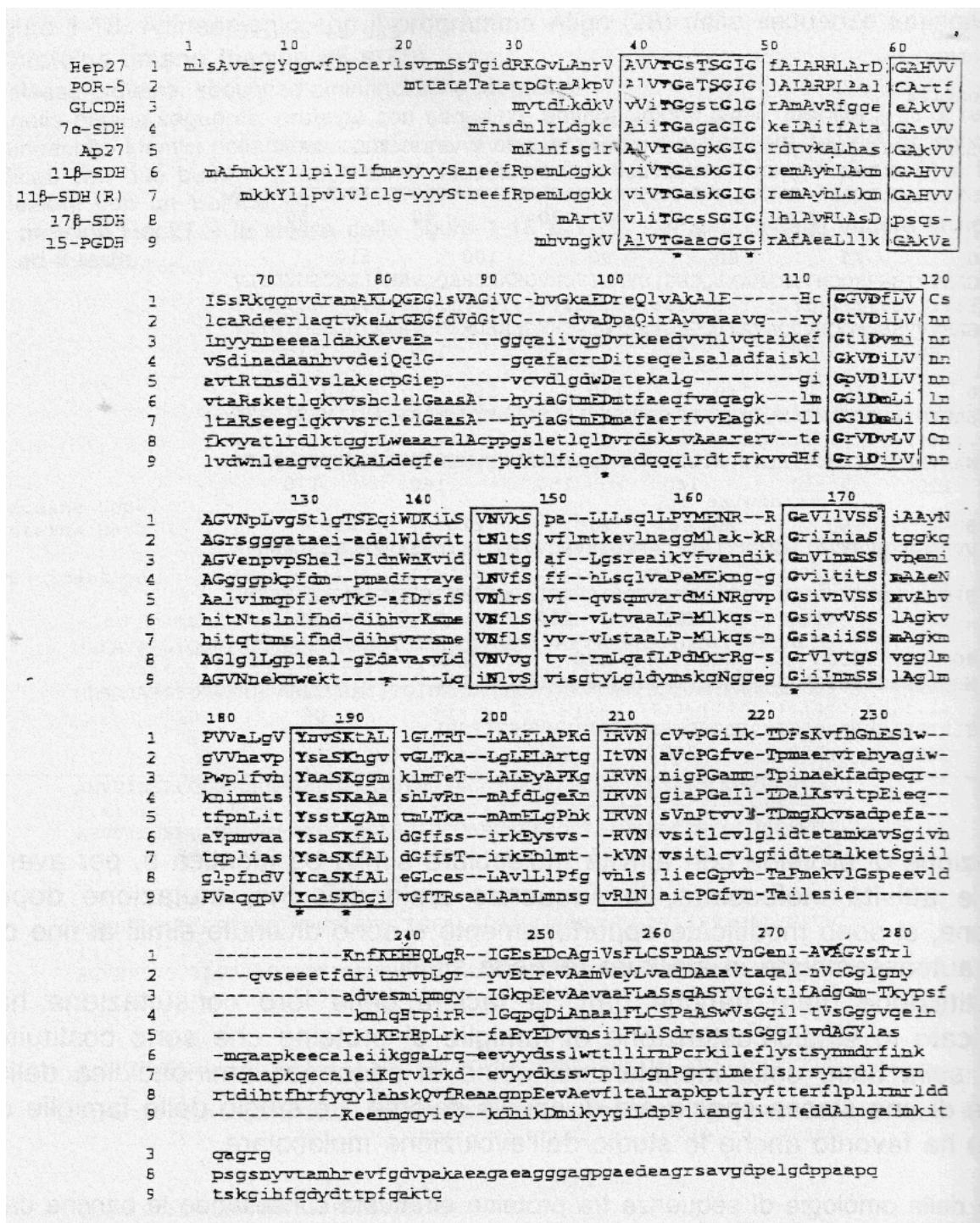



Figura 1-19. Allineamento multiplo di proteine appartenenti alla famiglia alcool-deidrogenasi a catena corta. Le sequenze simili sono indicate dai riquadri. Il dominio che lega il coenzima NAD (avvolgimento di Rossman) ha la sequenza compresa tra i residui aminoacidici 39-49. Esso include due glicine (Gly 43 e 47) separate da 3 aminoacidi. Le glicine avendo come residuo un atomo di H permettono un ripiegamento stretto del polipeptide. Il dominio che include i residui aminoacidici responsabili della catalisi (Tyr 185 e Lys 189) è posto tra i residui aminoacidici 185-192. Gli asterischi indicano i residui aminoacidici presenti in tutte le circa 1600 proteine di questa famiglia (residui fossili). Gli aminoacidi identici a quelli della prima sequenza sono scritti in maiuscolo ed anche in grassetto se identici nelle 9 proteine indicate in figura.

Ricerca delle caratteristiche e delle attività molecolari dei geni e delle proteine mediante analisi elettronica delle loro sequenze

La ricerca ed identificazione di sequenze consenso, di nota attività molecolare e funzione cellulare, nella sequenza nucleotidica di un gene e nella sequenza aminoacidica della proteina da esso codificata possono fornire informazioni sulle caratteristiche ed attività molecolari e sulle loro funzioni cellulari del gene e della proteina.
La ricerca delle sequenze consenso è effettuata elettronicamente utilizzando un computer collegato via internet a banche dati in cui sono conservate in archivi diversi le sequenze consenso di nota attività molecolare. Queste sequenze appartengono a gruppi di geni e di proteine (es. sequenza promotrice presente in geni diversi sensibile ad un dato fattore di trascrizione o dominio proteico presente in proteine diverse che associa un dato coenzima) e di ciascuna di esse è stata dimostrata con tecnologie chimiche e/o fisiche l'attività molecolare. Nell'archivio contenente tutte le sequenze dotate di una stessa attività molecolare è indicata anche la sequenza consenso costruita analizzando tutte le sequenze conservate nello stesso archivio. La sequenza del gene (o della proteina) di interesse è trascritta in apposite finestre dei programmi di gestione che poi vengono lanciati. La sequenza del gene viene confrontata mediante allineamento che simula una ibridazione, con tutte le sequenze dell'archivio elettronico scelto, es. l'archivio delle sequenze promotrici dei geni eucariotici. Al termine dell'elaborazione, sul monitor apparirà la sequenza consenso, sotto di essa la parte di sequenza del gene analizzato simile alla sequenza consenso ed ordinatamente per valore decrescente di similarità, le sequenze dell'archivio. In questo modo è possibile avere una buona valutazione del grado di similarità, tra la sequenza del gene di interesse analizzato e quelle già trovate in natura, per poi decidere se si può assumere che nella sequenza del gene di interesse ci sia la sequenza di nota funzione. Tuttavia anche nel migliore dei casi, 100% di identità di sequenza con una sequenza di nota attività molecolare, per avere la prova di quella attività occorre eseguire le analisi molecolari. Es. l'analisi dell'attività di una sequenza promotrice può essere fatta con la tecnologia del gene reporter (figura 2-2). L'identità di sequenza è una buona indicazione ma non è una prova effettuata sulla molecola di interesse. La sequenza consenso presente nel polipeptide può essere mascherata da altre parti della catena polipeptidica (esempio, essere all'interno della molecola).
Analogo procedimento è eseguito per analizzare le sequenze delle proteine: il programma di gestione è simile a quello per le sequenze di DNA e gli archivi includono le sequenze aminoacidiche delle proteine già identificate.

Analisi delle sequenze nucleotidiche dei geni

Sequenze dei promotori del gene. La presenza di queste sequenze può indicare la sensibilità/dipendenza dell'espressione del gene a particolari molecole. Ad esempio se tra i promotori del gene è presente una HRE (Hormone Responsive

Element), si ha l'indicazione che l’espressione del gene possa essere influenzata dall'azione di un ormone e quindi la sua azione può rientrare nel quadro delle azioni regolate dall'ormone e nelle relative patologie ormonali.

Esoni ed introni. Il migliore metodo per individuare la struttura di un gene in esoni ed introni è quella di allineare elettronicamente, con il programma Align di EXPASY o BLAST di NCBI, la sequenza del gene con quella del relativo cDNA (figura 1-20). La sequenza del cDNA verrà divisa in tante parti quanti sono gli esoni e le parti di sequenza si allineeranno sugli esoni lasciando libere e quindi evidenti le sequenze introniche. Le basi di confine degli introni ubbidiscono, salvo rarissime eccezioni, alla regola detta del GT-AG, cioè al 5' iniziano con GT ed al 3' terminano con AG. Questa regola permette di verificare quanto ottenuto con l'allineamento elettronico. Stabilire la struttura in esoni ed introni di un gene è utile per verificare se i cDNA sintetizzati da mRNA di cellule diverse sono il risultato di splicing alternativo. Con la clonazione del cDNA si può trovare cDNA diversi sintetizzati da mRNA presenti in cellule di tessuti diversi ed il confronto delle loro sequenze con quella del gene indica gli esoni che costituiscono i diversi cDNA e i diversi splicing. Nel caso che la sequenza del cDNA sia ancora ignota si può clonare utilizzando il DNA del gene come sonda e vagliando genoteche di più tessuti se non si conosce il tessuto di espressione del gene di interesse. Si possono anche usare specifici programmi di gestione che individuano la struttura in esoni ed introni di un gene basandosi sulle sequenze consenso di inizio della trascrizione e dei siti donatore /accettore dell'introne. Queste sequenze sono poste al 5' e al 3' del confine

```
----------------------------------------------------------------------------
1                                                                         76
AGCAAGCTGCTCTGGTTCAAATGCACGCTGTGGAAGCTTTGTTCTTTTGGTCTTCATGATAAATCTTGCTGCTGCT
                    ATGCACGCTGTGGAAGCTTTGTTCTTTTG             introne 1
    <------------------esone 1------------------>
77                                                                       152
CACTCGTTGGGTCCGTGCCACCTTACACACACACACACACACACACACAGGTCTGCAACTTCACTCCTGGGGCCAG
                        <----microsatellite---->
153                                                                      228
CAAGACCACGAATGCACCGAGAGGAATGAACAACTCTGGACACACCATCTTTAAGAACCGTAATACTCACCGCAAG
                     AGGAATGAACAACTCTGGACACACCATCTTTAAGAACC     introne 2
                     <---------------esone 2-------------->
229                                                                      304
GGTCTGCAACTTCATTCTTGAAGTCAGTGAGGCCAAGAACCCATCAATTCCGTACACATTTTGGTGACTTTGAAGA

305                                                                      380
GACTGTCACCTATCACCAAGTGGTGAGACTATTGCCAAGCAGTGAGACTATTGCCAATTGATGAGATCATCACCAA
                    TGGTGAGACTATTGCCAAGCAGTGAGACTATTGCCAAT
                    <---------------esone 3-------------->
381                                                                      456
GCGGTGAGACTATCACCTATCGCCAAGTGGCCTGATTCAGCAGGAAGCATCTCAGACACCAATAAACTAGCACAAC
                        esone 3 parte non tradotta
```

Figura 1-20. Allineamento con il programma Align (28) della sequenza di un cDNA con quella del rispettivo gene per individuare la struttura in esoni ed introni del gene. Sono sottolineati: il codice di inizio e di stop alla traduzione ed il segnale di poliadelinazione. I 5'GT ed i 3'AG degli introni sono in grassetto. L'introne 1 include un microsatellite AC.

tra introne ed esone. Un altro programma di gestione permette di individuare le sequenze degli esoni da quelle degli introni analizzando la composizione e le disposizioni relative delle basi. Le sequenze degli esoni, essendo sequenze codificanti, hanno particolari composizioni e disposizioni relative delle basi dovute alle restrizioni imposte dal codice genetico (le sequenze delle triplette hanno sequenze definite) mentre le sequenze degli introni, non essendo codificanti, non hanno quel tipo di restrizione.

Analisi delle sequenze aminoacidiche delle proteine

<u>Sequenze aminoacidiche che codificano domini aventi attività molecolare.</u>

Il dominio è una parte della struttura terziaria della proteina dotato di una specifica attività molecolare che partecipa all'attività molecolare della proteina. In genere un dominio è codificato da una sequenza consenso che può essere presente in proteine con attività molecolari simili, ma anche molto diverse. Comunque,l'identificazione del dominio fornisce informazioni sulla possibile attività molecolare della proteina.

Gli enzimi cinasi che associano ATP e trasferiscono un fosfato sul substrato, hanno un dominio, chiamato "ripiegamento di Rossman" (ripiegamento su se stessa della catena polipeptidica a livello di due glicine separate da 3 aminoacidi), che associa l'ATP interagendo prevalentemente con l'adenina-riboso. La diversità di attività molecolare delle cinasi è nei loro substrati (es. glucosio cinasi e 3P-gliceraldeide cinasi). Gli enzimi alcool deidrogenasi a catena corta hanno anch'essi un dominio di Rossman con il quale associano il coenzima NAD (Nicotinamide-Adenina-Dinucleotide) legando prevalentemente l'adenina-ribosio del NAD. La diversità di attività molecolare delle deidrogenasi è nei substrati ossidati o ridotti (17beta-testosterone deidrogenasi e 17beta-estradiolo deidrogenasi). I due domini degli enzimi cinasi e deidrogenasi hanno un certo grado di similarità ed assolvono ad una attività molecolare identica (legare un residuo di AMP) che si diversifica in relazione al coenzima legato ed all'azione specifica di altri domini che differenzia l'attività delle cinasi da quelle delle deidrogenasi.

Si ipotizza che i domini con la stessa attività molecolare possano essere omologhi, cioè originati da un dominio antenato comune, anche se parte di proteine non omologhe per la rimanente parte del polipeptide. Si assume che la sequenza del DNA che codifica un dominio proteico, nel tempo si sia duplicata più volte, poi le sue copie si sono inserite in altri geni ed hanno conferito alle proteine codificate da quei geni la nuova attività molecolare. Si ipotizza anche che l'origine di alcuni domini possa essere il risultato di una convergenza funzionale. I domini si sarebbero formati indipendentemente, e poiché dovevano assolvere ad una attività molecolare simile o identica (es. legare l'adenina-ribosio), mutazione dopo mutazione, sono evoluti fino ad avere sequenze simili per avere strutture simili capaci di una attività molecolare simile o identica.

Sequenze aminoacidiche segnale di localizzazione subcellulare.

Le sequenze segnale delle proteine possono essere poste in prossimità dell'aminoterminale (es. per la migrazione nel reticolo endoplasmatico e mitocondrio), del carbossiterminale (es. per la migrazione nei perossisomi), all'interno del polipeptide (es. per la migrazione nel nucleo, figura 1-16). I peptidi-segnale all'aminoterminale di alcune proteine, come quelli per la migrazione nel reticolo endoplasmatico e nel mitocondrio, sono eliminati quando la proteina ha raggiunto la localizzazione subcellulare definitiva. Questo spiega l'importanza di determinare il cDNA delle proteine, anche quando la loro sequenza proteica è già nota. Quando si purifica dalle cellule una proteina, essa è nella forma matura, pertanto facendo la sequenza aminoacidica della proteina purificata essa non includerà la sequenza del peptide segnale perché esso è stato precedentemente perso durante o dopo la migrazione della proteina. Mentre la sequenza della stessa proteina, dedotta dal cDNA, indicherà la presenza della sequenza del peptide-segnale.

Egualmente importante è conoscere la sequenza del gene, ed in particolare quella dei suoi esoni, per verificare la possibile presenza di splicing alternativo e quindi più forme di proteine generate dallo stesso gene, che in relazione al tipo di splicing possono includere o meno sequenze segnale.

La presenza nella sequenza aminoacidica di localizzazione nucleare (NTS, nuclear targeting sequence) indica che la proteina di interesse può essere localizzata nel nucleo pertanto l'attività molecolare della proteina e la proteina stessa deve essere ricercata nel nucleo. L'indicazione della localizzazione nucleare di una proteina è utile per procedere alla sua purificazione (si può iniziare a purificare la proteina partendo dai nuclei isolati piuttosto che da omogenati di cellule che includono anche le proteine citoplasmatiche) e per avere indicazioni sulla possibile attività molecolare e funzione fisiologica della proteina ricercata tra quelle compatibili con la fisiologia del nucleo. Tuttavia le proteine sono "magiche": una stessa proteina può essere localizzata in più di un distretto subcellulare (pur avendo la sequenza segnale per un singolo distretto cellulare) e svolgere in essi funzioni cellulari molto diverse: l'enzima citoplasmatico carbinolammina-deidratasi quando migra nel nucleo assume l'attività di fattore di trascrizione DCoH (appendice A); l'enzima mitocondriale aconitasi nel citoplasma è attivo come enzima ma ha anche la funzione di regolatore della traduzione del mRNA della ferritina (appendici A e B); la proteina Hep27, deidrogenasi a catena corta ha la sequenza segnale NTS per la localizzazione nucleare ma è anche localizzata nel citoplasma.

Sequenze aminoacidiche segnale per l'attacco di modificazioni post-traduzionali delle proteine.

Le proteine possono avere sequenze aminoacidiche riconosciute da:

-enzimi proteolitici per eliminare un peptide, ad es. per eliminare peptidi segnale all'amino- o al carbossi-terminale o per eliminare peptidi utili all'autoassemblaggio della proteina ma poi inutili o limitanti l'attività della proteina stessa ;

-enzimi che catalizzano reazioni covalenti su residui aminoacidici facenti parte della stessa sequenza, ad es. fosforilazione delle serine, treonine e tirosine, l'acetilazione e metilazione delle lisine ed arginine;
-enzimi che catalizzano la legatura covalente di glucidi o di polimeri glucidici alla glutamina della sequenza segnale .
La presenza in una proteina di uno o più siti di fosforilazione suggerisce che essa possa essere regolata da ormoni e/o fattori di crescita. I siti di glicosilazione (figura 1-16) suggeriscono la possibile glicosilazione della proteina che ne favorisce la migrazione in particolari distretti cellulari, es. lisosomi, membrane del reticolo citoplasmatico. L'acetilazione degli istoni suggerisce l'attività di decondensazione della cromatina associata ad attività di espressione genica (appendice B).

Sommario delle tecnologie di base per la sintesi, l'analisi e la ricerca dei geni

1 Digestione del DNA con enzimi di restrizione
2 Costruzione di mappe di restrizione
3 Reazione catalizzata dall'enzima ligasi
4 Analisi Southern
5 Sintesi di DNA mediante subclonazione
6 Tecnica della reazione a catena della DNA-polimerasi (PCR): sintesi ed analisi sequenza-specifica di DNA
7 Determinazione della sequenza nucleotidica del DNA
8 Conversione di mRNA in cDNA
9 Costruzione e vaglio delle genoteche
10 Programmi informatici per l'analisi dei dati delle sequenze nucleotidiche e amminoacidiche

Le analisi cliniche del fenotipo e del genotipo.

Prima dell'avvento delle tecnologie del DNA, la quasi totalità delle analisi cliniche dirette ad individuare le cause o i sintomi di uno stato patologico era effettuata esclusivamente sul fenotipo. Le analisi cliniche del fenotipo includono: l'analisi mediante microscopia o altri mezzi fisici (es. radiazioni elettromagnetiche) della morfologia degli organi, delle cellule e delle strutture subcellulari; l'analisi mediante mezzi fisici e biochimici delle caratteristiche molecolari delle cellule e dei fluidi extracellulari in particolare l'attività molecolare (es. catalisi enzimatica), la concentrazione delle proteine e dei metaboliti, la composizione di isoforme proteiche. Le analisi del genoma erano fatte mediante analisi citogenetiche con le quali si potevano osservare le alterazioni cromosomiche o indirettamente analizzando il fenotipo. Da pochi anni con l'introduzione delle tecnologie del DNA, si possono effettuare analisi direttamente su microquantità del genoma del paziente e su quello dell'agente patogeno che ha invaso il suo organismo. L'aspetto importante delle analisi genetico-molecolari è che esse possono dare indicazioni sulla predisposizione alla patologia anche molto tempo prima che essa si manifesti. Ciò è raramente possibile con le analisi morfologiche, fisiche e biochimiche del fenotipo perché se il fenotipo è cambiato (anche di poco) la patologia è già in corso ed essa può essere irreversibile. Le tecniche del DNA permettono di controllare lo stato (mutato o normale) dei geni che predispongono a patologie (anche di organi interni, analizzando il DNA di poche cellule del sangue evitando così le biopsie), di individuare portatori sani di patologie senza necessariamente conoscere gli ascendenti e/o i discendenti, di effettuare analisi prenatali di patologie geneticamente trasmesse o di patologie trasmesse da agenti patogeni prima che diventino numerosi e causino la malattia e di effettuare analisi di patologie di cui non si conosce ancora la proteina o la proteina è nota ma è di difficile dosaggio.

Quelli che s'innamoran di pratica senza scienza, son come 'l nocchiere ch'entra in navilio senza timone e bussola che mai ha certezza dove si vada. *Leonardo da Vinci*

Capitolo 2.

Tecnologie per ricercare la funzione dei geni

Il completamento del progetto genoma umano ha permesso di conoscere molti geni di proteine che erano sconosciute e che sarebbe stato molto difficile individuare con metodi biochimici perché espresse in piccole quantità, in un ristretto numero di cellule o perché molto labili. Determinata la sequenza del gene è tecnologicamente molto facile determinare la sequenza della proteina codificata ma occorre una ricerca accurata e spesso non facile per arrivare a definire l'attività molecolare e la funzione fisiologica della proteina. Questa ricerca è detta biochimica inversa perché procede in senso inverso rispetto alla biochimica tradizionale. Nella biochimica tradizionale la determinazione della sequenza aminoacidica di una proteina era l'ultimo atto di una spesso lunga ricerca iniziata con l'aver individuato rispettivamente in cellule umane normali una attività molecolare o una funzione cellulare ed in cellule malate la perdita di attività e funzioni molecolari. La ricerca continuava con determinazione fino alla purificazione della proteina utilizzando l'attività molecolare della proteina come segnale della presenza della proteina durante le procedure di purificazione (vedere clonazione funzionale dei geni, capitolo 4). Avere la proteina pura permetteva poi di intraprendere il complesso e lungo lavoro della determinazione dell'intera sequenza aminoacidica della proteina.
Dagli anni '70 in poi sono state messe in funzione tecnologie che permettono di ricercare la funzione del gene che è la funzione cellulare della proteina da esso codificata. L'importanza di queste tecnologie è nel poter utilizzare il DNA del gene per eseguire esperimenti che non sono possibili o sono comunque molto più laboriosi se eseguiti utilizzando la proteina codificata dallo stesso gene (Tabella 2-1).
Avendo la disponibilità della molecola di un gene o del suo cDNA, le biotecnologie molecolari e cellulari permettono di sintetizzare *in vitro* la proteina codificata al fine di caratterizzare la sua molecola e la sua funzione cellulare, di individuare l'organo/tessuto dove la proteina è sintetizzata e la sua localizzazione subcellulare. Ed inoltre, di valutare l'entità dell'espressione del gene in differenti condizioni metaboliche, durante lo sviluppo embrionale, l'organogenesi e il ciclo cellulare, di mutare il gene stesso, di incrementare, modulare od annullare la sua espressione e di conseguenza alterare, incrementare, modulare od annullare l'attività molecolare della proteina codificata e monitorare gli effetti provocati sulle funzioni cellulari. La funzione del gene è poi dedotta dalle conseguenti alterazioni della fisiologia della cellula e/o dell'organismo. Queste stesse alterazioni fenotipiche danno anche indicazioni su possibili manifestazioni patologiche conseguenti all'alterazione della struttura o della regolazione del gene.
Le più comuni biotecnologie sono schematizzate di seguito.

Tecnologia dell'espressione del cDNA in cellule batteriche o in cellule eucariotiche per produrre la proteina di interesse.
L'inserimento del cDNA legato al vettore privo di proteine (nudo) in cellule batteriche è detto trasformazione (figura 1-7b) ed in cellule eucariotiche è detto transfezione.
La transfezione in cellule batteriche è effettuata mescolando il costrutto cDNA-vettore con le cellule e scaldando per un breve tempo (vedere subclonazione e figura 1-7). La transfezione di cDNA di mammifero è più laboriosa, il costrutto cDNA-vettore è aggiunto alla sospensione di cellule e, la sua penetrazione all'interno di esse, viene favorita dalla co-precipitazione sulla membrana plasmatica del cDNA-vettore con i sali di Calcio, dall'associazione del cDNA-vettore con liposomi (vescicole lipidiche artificiali) che si fondono con la membrana plasmatica oppure mediante elettroporazione (figura 2-4b). Con uno dei metodi sopra indicati, il cDNA del gene di interesse viene legato ad un plasmide di espressione e transfettato nelle cellule per produrre grandi quantità della proteina codificata al fine di analizzare la sua struttura terziaria e quaternaria ed i residui aminoacidici posti sulla superficie della proteina (residui che conferiscono alla proteina la possibilità di associare farmaci, inquinanti vedere figura D-7), l'attività molecolare (es. verificare se è un enzima con attività catalitica su un dato substrato), e per produrre anticorpi contro la proteina da utilizzare per analisi Western. Gli anticorpi sono utili per valutare la presenza di piccolissime quantità di proteina nelle cellule, distretti subcellulari ed extracellulari normali o patologici di uno stesso organismo ed avere indicazioni sulla funzione della proteina deducendole dall'espressione differenziata del gene (es. proteina presente esclusivamente nelle ghiandole surrenali). La transfezione è utilizzata anche per scopi industriali. L'insulina ed il fattore di coagulazione VIII utilizzati in terapia sono prodotti dalle industrie utilizzando cellule geneticamente modificate mediante transfezione.
Transfettare il cDNA, piuttosto che l'intero gene, semplifica la tecnologia perché sono evitate le reazioni della maturazione del mRNA, ciò è particolarmente importante nelle transfezioni in cellule batteriche che non posseggono le proteine necessarie. La transfezione del cDNA nelle cellule eucariotiche di lievito è preferita quando la proteina espressa dal gene studiato subisce modificazioni covalenti (es. fosforilazione, glicosilazione) necessaria alla sua attività molecolare, modificazioni non operate o operate diversamente dalle cellule batteriche.

Tecnologia della transfezione del gene in cellule omologhe in cultura. La transfezione del DNA di un gene umano in cellule omologhe in cultura può avvenire come frammento di DNA genomico nudo o legato ad un plasmide nudo e transfettato nelle cellule con uno dei metodi sopra indicati. Quando il vettore è un virus animale il DNA del gene penetra nelle cellule impacchettato nella capsula virale e la transfezione è detta transduzione. Il DNA transfettatto raggiunge il nucleo della cellula ed in genere non è integrato nel DNA nucleare rimanendo come episoma (DNA extracromosomico). Se il DNA del gene è

legato ad un vettore che ha una origine di replicazione si formano molte copie del costrutto ed esse rimangono nelle cellule al massimo per 2-3 giorni; se il vettore non ha l'origine di replicazione la permanenza del DNA esogeno nella cellula è di poche ore. Raramente, il DNA del gene viene integrato permanentemente nel DNA di un cromosoma della cellula ospite (transgene integrato), l'integrazione è detta trasformazione e la cellula eucariotica è detta trasformata perché ha il fenotipo cambiato. L'integrazione avviene dove un filamento del DNAds cromosomico è interrotto per la mancanza di un legame fosfodiestere. Il senso originale del termine trasformazione era per le cellule che trasformate replicavano indefinitamente prendendo il nome di linea cellulare (notare che la trasformazione nei batteri ha un altro significato).
Il DNA del gene può essere combinato con un promotore esogeno particolare che è attivato da piccole molecole o ioni. Dopo transfezione le cellule sono esposte al legante-segnale al fine di vedere gli effetti del prodotto genico sulla fisiologia cellulare. Il promotore sensibile a leganti-segnale esogeni è necessario quando si vuol verificare l'attività di geni che controllano negativamente la replicazione cellulare. Dopo qualsiasi transfezione occorre far replicare le cellule alcune volte, affinché recuperino dalle alterazioni provocate dalla transfezione (es. sulla membrana plasmatica) e per poter verificare se la transfezione è realmente avvenuta. Se un gene, inibitore della replicazione, fosse attivo appena transfettato, le cellule verrebbero inibite nella replicazione e, non potendo recuperare dallo shock della transfezione, morirebbero rendendo impossibile condurre l'esperimento. Cioè, verificare se l'arresto delle crescita cellulare sia dipendente dalla proteina transfettata o dai danni della transfezione.
L'espressione del gene nella cellula ospite causa un eccesso di concentrazione della proteina codificata che a sua volta causa un cambiamento della fisiologia della cellula e da esso si possono avere indicazioni sulla funzione e talvolta anche sull'attività molecolare della proteina espressa.
In alcuni casi, dopo transfezione di un cDNA, il fenotipo cellulare (morfologia e funzione) può rimanere invariato perché i meccanismi di regolazione della cellula possono inibire l'azione del cDNA.

Tecnologia della transfezione del gene in cellule omologhe mutate. Con questa tecnologia, il gene viene transfettato in cellule omologhe mutate mancanti di una attività molecolare. La transfezione ha lo scopo di verificare che l'espressione del gene transfettato restauri l'attività mancante nelle cellule riceventi. In caso positivo, si ha l'indicazione che il gene transfettato codifica la proteina mancante o inattiva nelle cellule perché mutata. Da ciò si ottiene l'identificazione e la clonazione (purificazione) del gene mutato. Questo tipo di transfezione è utilizzato anche per individuare e clonare geni della suscettibilità a patologie con il saggio di complementazione (capitolo 1).
La transfezione di un gene normale in cellule mutate patologiche è il principio su cui si basa la terapia genica, che ha lo scopo di restaurare le funzioni cellulari perse per causa di mutazioni geniche.

Tecnologie per la transfezione del gene in animali. Per ottenere animali transgenici occorre che il DNA del gene di interesse si integri nel DNA nucleare dell'ospite e sia presente in tutte le sue cellule, pertanto la transfezione viene operata su zigoti o giovani embrioni. L'integrazione del DNA avviene in punti in cui casualmente manca il legame fosfodiestere tra due nucleotidi. Il DNA del gene può essere transfettato mediante microiniezione nel pronucleo maschile di topo (Mus musculus) (figura 2-1a) oppure in cellule prelevate dalla massa centrale della blastocisti di topo mediante elettroporazione (figura 4-4b) o veicolato da retrovirus. Per ricerche sulla funzione di geni umani in genere viene utilizzato il topo, tuttavia con la stessa tecnica sono stati utilizzati anche altri animali. Le cellule prelevate dalla blastocisti sono cellule staminali embrionali (ES cells capaci di differenziarsi in vari tessuti) che poi vengono reinserite in un'altra blastocisti e partecipano alla formazione dell'embrione (figura 2-4).
Con la transfezione nel pronucleo maschile è più probabile che tutte le cellule che si formano dallo zigote abbiano integrato il gene di interesse, quindi anche le cellule germinali, e abbiamo così la possibilità di produrre altri soggetti transgenici mediante accoppiamenti. Talvolta il DNA del gene transfettato si integra con il DNA genomico dopo alcune divisioni dello zigote e il topo neonato è un mosaico di cellule transfettate e non transfettate. Con la transfezione in cellule della blastocisti, l'animale è una chimera di cellule transfettate e cellule normali provenienti rispettivamente da zigoti diversi. Quindi occorre fare più tentativi di transfezioni fino ad avere un maschio ed una femmina con cellule germinali portatrici del gene transfettato. Quando i topolini neonati sono un mosaico (transfezione in pronuclei), chimerici (transfezione in blastocisti) e eterozigoti occorre procedere a reincroci (genitori con figli e tra sorelle e fratelli) al fine di ottenere almeno un maschio ed una femmina con il gene transfettato in tutte le loro cellule e, quando è utile per la ricerca si portano ad essere anche omozigoti.
Dall'analisi degli effetti sul fenotipo causati dall'espressione del gene durante lo sviluppo embrionale e nell'adulto si hanno indicazioni sulla sua funzione (figura 2-1b). L'espressione genica è maggiore perché l'animale esprime il proprio gene omologo a quello transfettato che talvolta è presente in più copie in tandem nel punto di integrazione. Il gene transfettato può essere munito di un promotore/enhancer tessuto specifico e ciò determina l'espressione del gene transfettato nel tessuto di interesse. Il gene transfettato può essere omologo (proveniente dalla stessa specie), eterologo o essere uno mutato.
Il gene mutato da indicazioni sulla sua capacità a causare patologie, se la mutazione porta ad un prodotto genico con eccesso di attività o con nuove proprietà.
La transfezione di un oncogene umano con un attivo promotore fegato-specifico in blastocisti di topi ha permesso di osservare nei topi, da essi discendenti, la formazione di tumori epatici alcuni mesi dopo la nascita. Topi neonati portatori di due diversi oncogeni umani, ottenuti da incroci dei precedenti, anticipavano l'insorgenza dei tumori ad un mese dopo la nascita.

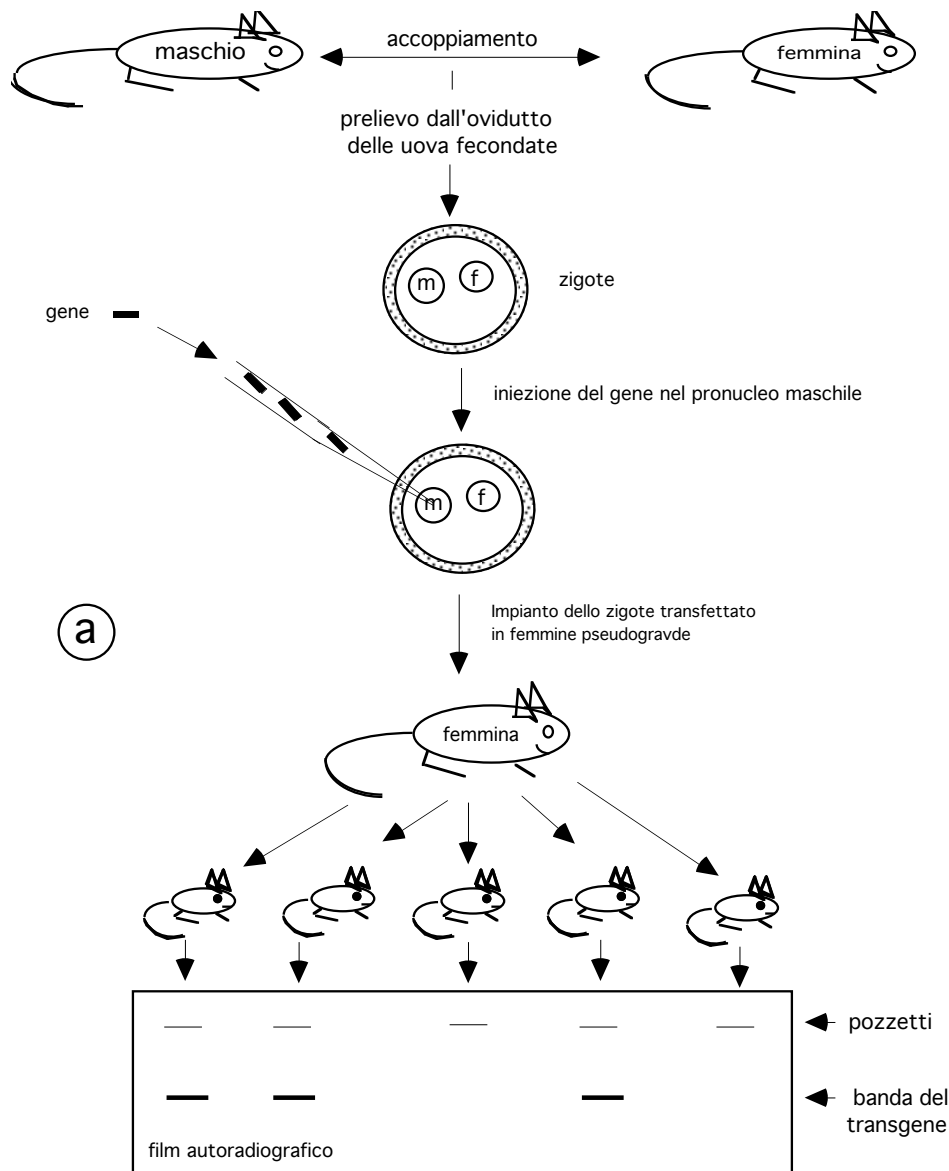



Figura 2-1a) Tecnologia per la produzione dell'animale transgenico mediante inserimento del gene nel pronucleo maschile. Uova fecondate (zigoti) sono prelevate dall'ovidotto di femmine di topo (Mus musculus) ed il gene di interesse è microiniettato con speciali sottilissime micropipette di vetro in uno dei due pronuclei (in genere quello maschile perché è più grosso). Il volume iniettato è circa 2 picolitri e contiene circa 2-3 centinaia di copie del gene. Gli zigoti transfettati sono impiantati di nuovo in femmine pseudogravide, rese tali mediante accoppiamento con maschi resi sterili mediante vasectomia. Dopo circa tre settimane, nei neonati topini la presenza del gene è verificata sul film autoradiografico della tecnica Southern o sul gel di agarosio di una PCR analitica. In figura da sinistra il 1°, 2° e 4° topo risultano positivi. Il transgene è integrato a caso nel DNA cromosomico in punti in cui manca un legame fosfodiestere (nick = tacca)) e l'integrazione può includere più copie in tandem nello stesso locus. L'integrazione è stabile ed interessa tutte le cellule (cellule somatiche e germinali). Il gene che può essere un gene normale omologo o eterologo, mutato o reporter, è trasmesso nella discendenza come i geni naturali. La discendenza è in genere eterozigote e mediante incroci si possono ottenere individui omozigote. (ridisegnato e modificato da Recombinant DNA, Watson J.D., Gilman M., Witkowski J., Zoller M. (1992) Scientific American Books, 2nd ed., e da Strachan T. and Read A.P. (1996). Human Molecular Genetics. Bios, UK).

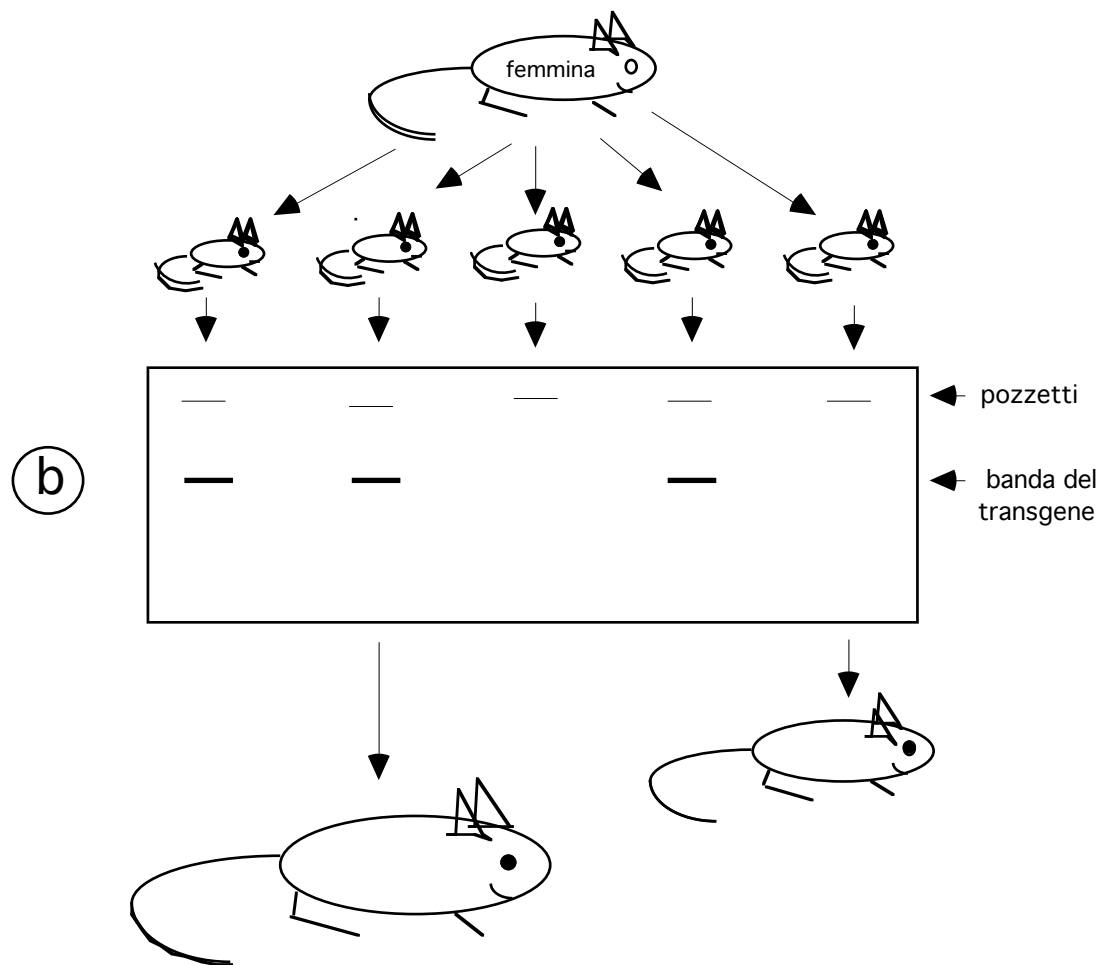



Figura 2-1. b) Con la tecnica indicata in a) è stato transfettato in topi (Mus musculus) il gene dell'ormone della crescita. I topi transgenici raggiungono, rispetto ai topi di controllo, peso e dimensioni maggiori. La tecnologia permette di osservare la funzione fisiologica del gene transfettato: stimolazione della crescita mantenendo normali le proporzioni e le posizioni relative degli organi (ridisegnato e modificato da Recombinant DNA, Watson J.D., Gilman M., Witkowski J., Zoller M. (1992) Scientific American Books, 2nd ed., e da Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK).

Si dimostrava così la cooperatività degli oncogeni nella cancerogenesi e la natura multigenica dei tumori.

Il gene transfettato in relazione alla posizione con la quale è inserito nel DNA cromosomico può avere una espressione quantitativamente o temporalmente diversa dal gene endogeno ad esso omologo e causare alterazioni della normale fisiologia dell'organismo. Si assume che lo stato della cromatina, cioè delle proteine associate alla regione cromosomica in cui casualmente si è inserito il gene, ne influenzi l'espressione. La diversa espressione del gene transfettato è detta ectopica (posizione diversa).

Tecnologia del gene reporter. I geni reporter sono geni chimerici (DNA da due geni diversi) realizzati mediante ricombinazione del DNA della regione dei promotori del gene del quale vogliamo conoscere la regolazione e dal DNA della parte codificante di un gene che rende visibile l'espressione del gene e per questo è detto gene reporter. Il gene reporter codifica una proteina, non presente nelle cellule ospiti, il cui dosaggio quantitativo è semplice da eseguire (figura 2-2a).
Come geni reporter sono stati utilizzati geni codificanti:
- l'enzima CAT = cloroamfenicolo acetil-transferasi che utilizza come substrato il cloroamfenicolo C^{14} (radioattivo) che è evidenziato mediante autoradiografia.
- l'enzima beta-galattosidasi batterico che con il substrato Xgal forma un precipitato blu.
- l'enzima luciferasi che catalizza la reazione di ossidazione della luciferina con emissione di luce giallo-verde,
- proteine di meduse che naturalmente sono fluorescenti verdi.
I geni reporter transfettati in cellule in cultura omologhe permettono di verificare se il gene è espresso in una particolare fase del ciclo cellulare, se la sua espressione è sensibile a fattori di regolazione endogeni, prodotti dall'organismo umano (ormoni, vitamine, metaboliti, ioni) e a molecole esogene (farmaci, anestetici, allergeni, inquinanti ed altre molecole esogene). La stessa tecnica è utilizzata per ricercare le sequenze promotrici del gene di interesse e verificare la loro tessuto specificità (figura 2-2b). La regione promotrice del gene di interesse viene modificata, base per base su tutta la sua lunghezza, mediante delezioni o mutazioni puntiformi ambedue sito-specifiche (figura 2-3) e legata covalentemente al gene reporter. I vari costrutti chimerici sono transfettati in gruppi diversi di cellule in cultura ed in questo modo si individuano le sequenze che controllano l'espressione del gene e se il gene è transfettato in giovani embrioni (zigoti o blastocisti) è possibile individuare le sequenze che conferiscono tessuto-specificità.
Il gene reporter transfettato in pronuclei (figura 2-1a) o in cellule di blastocisti di topo (figura 2-4), permette di osservare l'espressione del gene durante lo sviluppo embrionale fino all'adulto e definire in quali stadi embrionali, fetali e post-natali (adulto incluso) ed in quali tipi di tessuti/organi sia espresso il gene di interesse (figura 2-2b).
La stessa tecnica permette di verificare l'espressione del gene in un tessuto/organo posto in un dato stato metabolico (es. digiuno, alimentato con

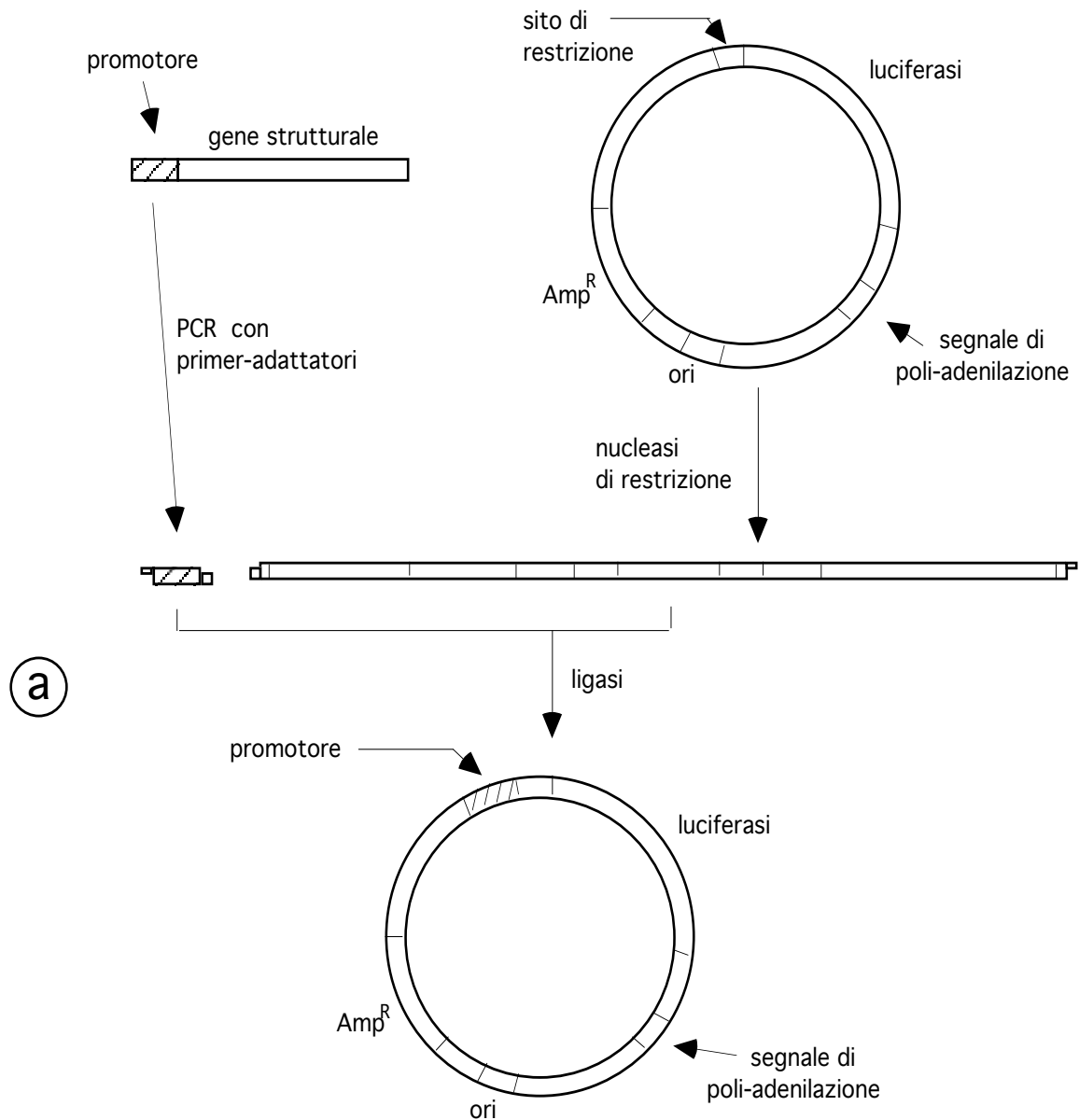



Figura 2-2. a). La regione del promotore del gene del quale vogliamo valutare l'espressione viene amplificata con la PCR utilizzando due primer che oltre alla sequenza di associazione al DNA del promotore hanno al loro 5' un tratto di DNA che include una sequenza di restrizione. L'amplificato viene digerito con un enzima di restrizione e legato con ligasi al plasmide digerito con lo stesso enzima di restrizione. Il promotore di interesse è legato al gene reporter della luciferasi completo di segnale di poliadenilazione. Il costrutto è in genere linearizzato con un taglio operato da un enzima di restrizione per favorire l'integrazione nel DNA nucleare e poi transfettato in cellule in cultura, in zigoti o in cellule di blastocisti.
(ridisegnato e modificato da Strachan T. and Read A.P. (1999) Human Molecular Genetics, Bios, UK).

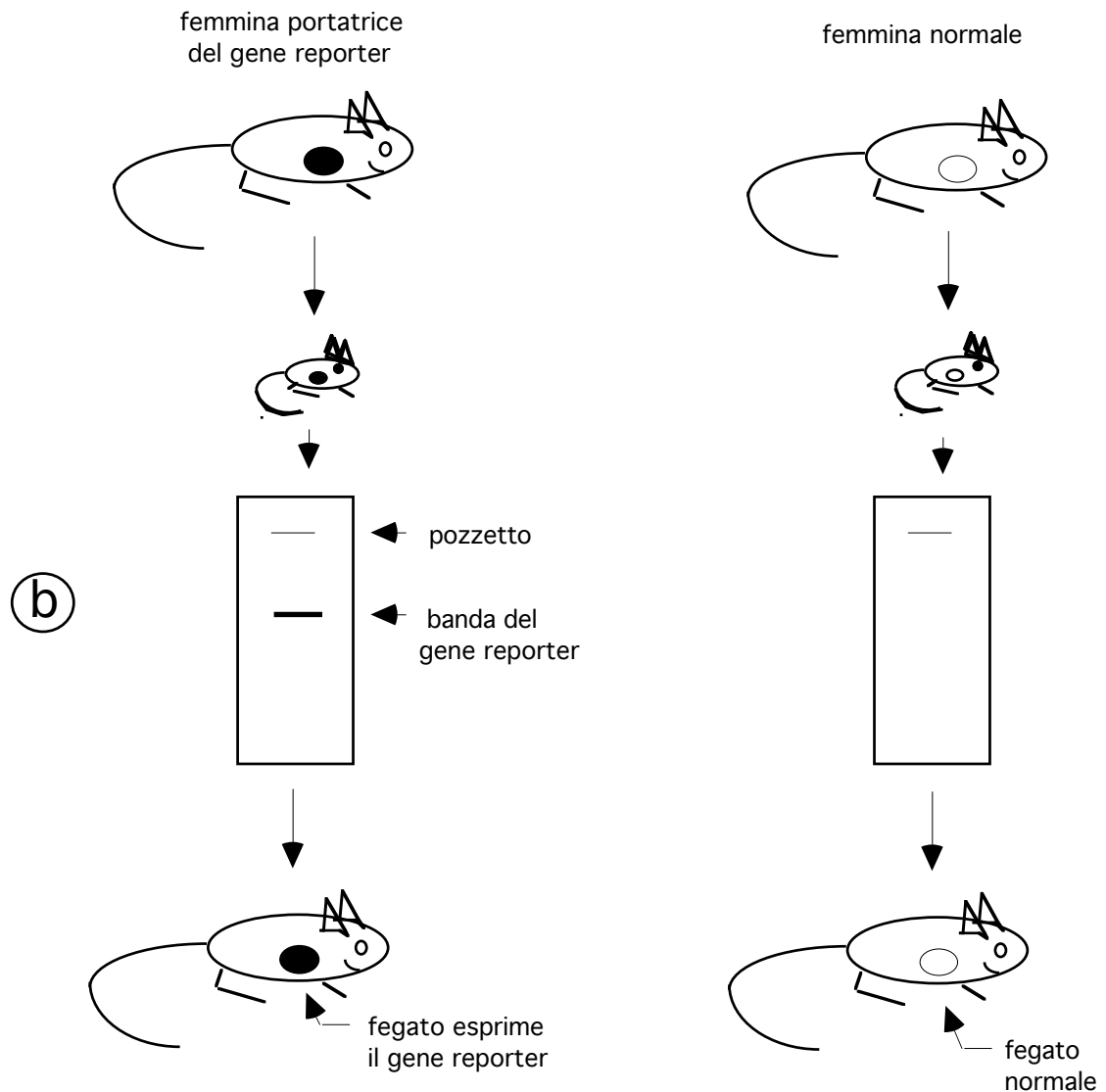



Figura 2-2. b). Un costrutto simile a quello indicato in figura 2-2a viene inserito in tutte le cellule di in una femmina di topo (Mus musculus) mediante transfezione in zigoti (2-1 a) o in cellule di blastocisti (2-4b) e successivi reincroci. L'analisi Southern o PCR analitica in qualsiasi cellula dei topi neonati mostrerà la presenza del gene reporter (topo a sinistra). Un analisi mediante luciferina condotta in tutti i tessuti mostra che la proteina reporter (luciferasi) è presente esclusivamente nel fegato adulto e solo negli epatociti del parenchima (non in altre cellule dello stesso organo). La tecnologia ha permesso di osservare che quel gene ha un promotore fegato specifico.

una particolare dieta) o esposto a fattori di regolazione endogeni o a molecole esogene ed inoltre in organi in definiti stati patologici. Per eseguire questi esperimenti occorre sacrificare l'animale adulto o l'embrione nei vari stadi embrionali ed esporre fettine del loro tessuti (da osservare poi al microscopio ottico) al substrato dell'enzima reporter o alla luce di una data lunghezza d'onda per le proteine fluorescenti (in questo caso si usa il microscopio a fluorescenza). La quantità del prodotto di reazione radioattivo, colorato, la quantità di luce emessa durante la reazione luciferina-luciferasi o della luce emessa dalla proteina fluorescente opportunamente eccitata rivelano la quantità della proteina del gene reporter.

Mutazione sito-specifica del gene. Il DNA di un gene o il suo cDNA possono essere mutati *in vitro* e poi transfettati in cellule batteriche, eucariotiche o in zigoti e blastocisti. Le mutazioni, che possono essere puntiformi, permettono di identificare e di definire esattamente con la risoluzione di una sola base, le sequenze promotrici, le sequenze segnale per lo splicing del mRNA, le sequenze segnale di inizio e di stop alla trascrizione ed inoltre permettono di avere informazioni sul ruolo che queste sequenze hanno nella regolazione dell'espressione genica. La mutazione che causa una riduzione dell'espressione genica dimostra che la sequenza modificata ha un ruolo nel controllo dell'espressione genica.
La mutazione sito-specifica della regione promotrice di un gene è utilizzata anche per individuare tessuto specificità delle singole sequenze promotrici e la loro sensibilità ai fattori di trascrizione. Mutazioni delle parti codificanti dei geni (esoni) sono operate su geni, omologhi a quelli umani, appartenenti a cellule di blastocisti di animali al fine di verificare se nell'animale adulto una data mutazione di un dato gene, sia quella responsabile di una data patologia osservata nell'uomo.
Le mutazioni operate su i cDNA, che poi sono transfettati e tradotti *in vitro*, permettono di avere informazioni sul ruolo che singoli domini proteici hanno nell'attività molecolare, nella regolazione da effettori e sulla funzione cellulare della proteina codificata. La stessa tecnologia permette di caratterizzare il ruolo che singoli residui aminoacidici hanno nell'attività molecolare e nel mantenimento della struttura della proteina.
Per analizzare gli effetti di mutazioni di regioni del promotore è utilizzata la tecnica del gene reporter (figura 2-2); per analizzare il ruolo delle sequenze segnale dello splicing del mRNA, di inizio o di segnale di stop della trascrizione si analizza la sequenza degli mRNA prodotti in vitro. Per analizzare il ruolo dei domini o dei residui aminoacidici delle proteine occorre possedere un metodo di dosaggio dell'attività molecolare della proteina stessa (es. per gli enzimi, il dosaggio dell'attività catalitica, per i recettori degli ormoni la valutazione della capacità ad associare il relativo ormone).

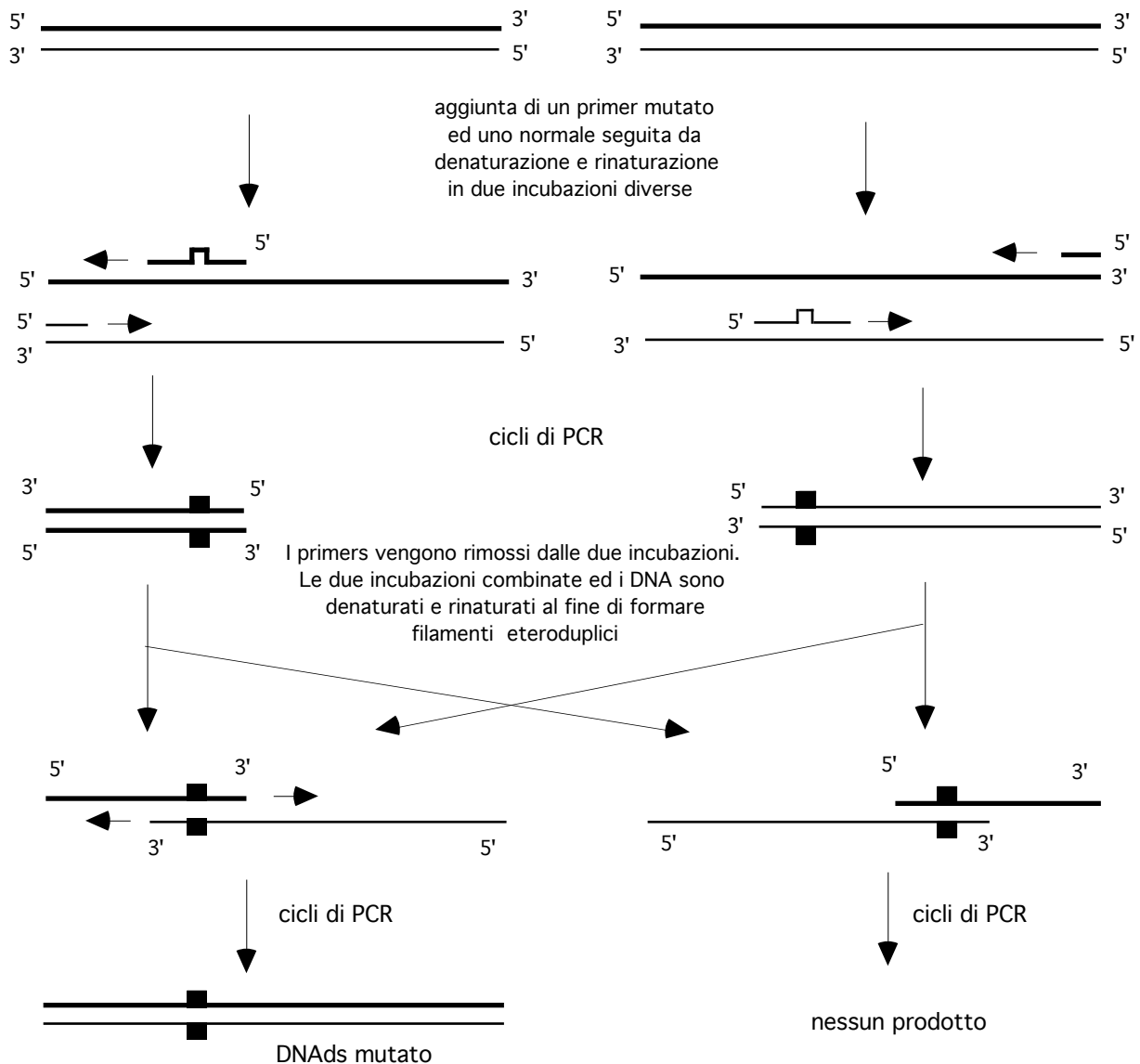



Figura 2-3. Mutazione sito specifica realizzata con PCR. Per realizzare la mutazione si sintetizza una coppia di primer complementari alla regione di DNA contenente la base da mutare. I due primer sono mutati nella base (sporgenza rettangolare) che si vuol sostituire alla base normale e esattamente complementari tra loro. Si preparano due miscele di incubazione, ciascuna con un primer mutato ed uno normale e si operano i cicli di PCR. Nelle due incubazioni verranno amplificate delle regioni diverse del DNA stampo in parte sovrapponibili nella regione di associazione dei primer. In figura la coppia di basi mutate divenute complementari dopo i cicli di PCR è indicata con due rettangolini neri.
Successivamente vengono eliminati i primer dalle miscele di reazione e le due soluzioni combinate e poi sottoposte a temperature di denaturazione e rinaturazione per formare casualmente i filamenti DNAds eteroduplici. Casualmente si riformeranno anche i filamenti di DNAds non in forma eteroduplice così come erano stati precedentemente sintetizzati (non sono mostrati in figura). Si sottopone la soluzione a nuovi cicli di PCR. Un DNAds eteroduplice permetterà di pro durre copie di filamento stampo mutate nella posizione e base di interesse, l'altro non può essere amplificato dato l'orientamento dei due filamenti di DNAss e così anche i filamenti associati non in forma eteroduplice. In figura per comodità di disegno, i filamenti di DNAds sono indicati con spessori diversi (ridisegnato e modificato da Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK).

Le proteine da analizzare sono sintetizzate *in vitro* mediante transfezione in cellule batteriche o eucariotiche di cDNA piuttosto che per transfezione di geni perché la manipolazione del cDNA (operare su esso mutazioni, transfettarlo ed ottenere *in vitro* la relativa proteina) è più semplice di quella dei geni.
La tecnica della mutazione sito specifica ha rivoluzionato gli studi di genetica tradizionale perché permette di progettare ed operare mutazioni anche di singole basi in posizioni specifiche della sequenza nucleotidica del gene di interesse.
Prima della messa a punto di questa tecnologia occorreva ricercare/analizzare il fenotipo che era stato mutato a caso dagli eventi naturali o sperimentalmente con radiazioni o reagenti.
Il procedere di questa tecnologia: "induzione della mutazione sito specifica ---> analisi del fenotipo modificato" risulta inverso rispetto a quello della genetica tradizionale "analisi del fenotipo alterato ---> assunzione che il gene sia mutato", indusse a chiamarla "genetica inversa" per distinguerla dalla genetica tradizionale. In realtà la tecnologia non ha niente di inverso perché è diretta proprio dal gene al fenotipo ed ora si evita di chiamarla genetica inversa.
In figura 2-3 è indicato il metodo di operare la mutazione sito-specifica utilizzando la PCR, esiste un metodo simile che per ottenere frammenti di DNA mutati, utilizza anch'esso dei primer mutati per replicare il DNA di interesse inserito in un plasmide.

Tecnologie per l'inibizione dell'espressione di singoli geni

Tecnologia per la distruzione di geni in animali (gene knockout).
Il gene di cellule di blastocisti, prelevate da una femmina di topo (Mus musculus), viene distrutto mediante ricombinazione omologa doppia con un frammento di DNA mutato. Per attuare la tecnologia di distruzione del gene mostrata in figura 2-4 (ne esistono anche altre) viene costruito un plasmide che include delle regioni di DNA identiche a quelle del gene di interesse per poter operare la ricombinazione omologa doppia. All'interno di queste sequenze sono inseriti dei codoni di stop o tolte sequenze importanti al fine di rendere il gene non più capace di codificare una proteina attiva o di non sintetizzare più lo mRNA. Inoltre viene inserito un DNA marcatore che è utilizzato per verificare mediante PCR se è avvenuta la ricombinazione e quindi la distruzione del gene. L'inserimento del DNA di distruzione del gene nel plasmide permette di produrre per subclonazione molte copie del costrutto e poi di poter operare la distruzione del gene in molte cellule di blastocisti. Per favorire la ricombinazione il plasmide deve essere tagliato con un enzima di restrizione in un singolo sito (linearizzazione) esterno alla regione di ricombinazione ed in questa forma il costrutto è detto vettore per colpire il bersaglio (targeting vector), cioè il gene che vogliamo inattivare. Il vettore per colpire il bersaglio è transfettato mediante elettroporazione (figura 2-4b). Con la stessa modalità si può determinare la mutazione di una singola base del gene di interesse, operando

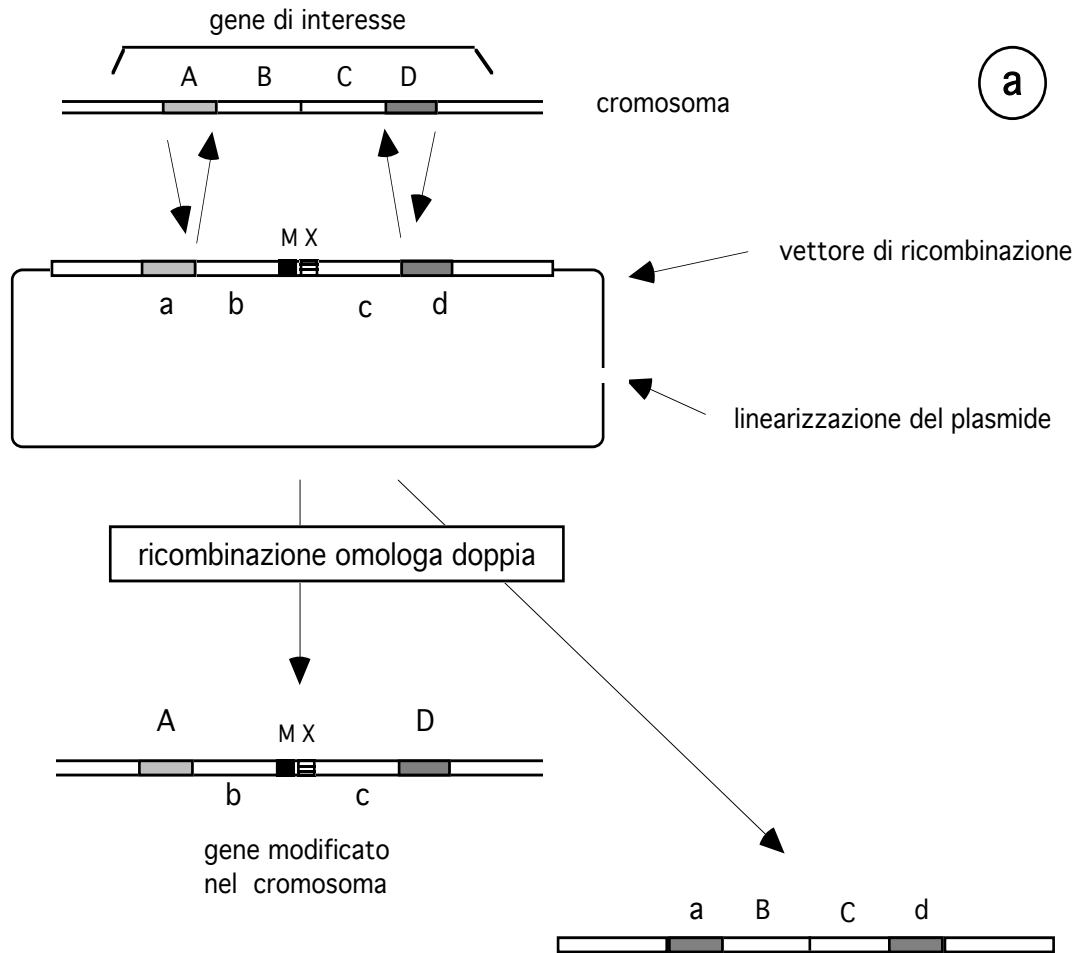



Figura 2-4. a). Distruzione del gene mediante ricombinazione omologa doppia. Viene costruito un plasmide che include regioni di DNA (a, b, c, d) aventi parti con sequenze identiche a quelle (A, B, C, D) del gene di interesse per poter operare la ricombinazione. Le regioni cromosomiche A-B e C-D si associano rispettivamente alle regioni del costrutto a-b e c-d e la ricombinazione avviene tra A e B, a e b, C e D, c e d. All'interno di queste sequenze sono operate mutazioni (X) ed inserito un marcatore (M) che è utilizzato per selezionare le cellule in cui è avvenuta la ricombinazione e quindi la distruzione del gene. Per operare la ricombinazione il plasmide deve essere tagliato con un enzima di restrizione in un singolo sito (linearizzazione) divenendo il vettore (targeting vector) per colpire il bersaglio costituito dal gene che vogliamo inattivare. (ridisegnato e modificato da Strachan T. and Read A.P. (1996) Human Molecular Genetics, Bios, UK).

Figura 2-4. b) Pagina successiva. Produzione di un topo (Mus musculus) mutato specificamente in un gene. In relazione al costrutto sintetizzato, la mutazione può causare la distruzione del gene (figura 2-4a) od interessare una singola base. Inizialmente si ha la nascita di topi chimerici, parte delle loro cellule provengono dalle cellule mutate mediante ricombinazione omologa (indicate in nero) e parte sono normali. Quando si ottengono topi con cellule germinali mutate, mediante reincroci è possibile ottenere individui costituiti completamente da cellule mutate. I disegni sono indicativi, le dimensioni relative non sono rispettate ed in particolare la cella per l'elettroporazione ha forma ed elettrodi diversi.
(ridisegnato e modificato da Recombinant DNA, Watson J.D., Gilman M., Witkowski J., Zoller M. Scientific American Books, 2nd ed., (1992) e da Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK.)

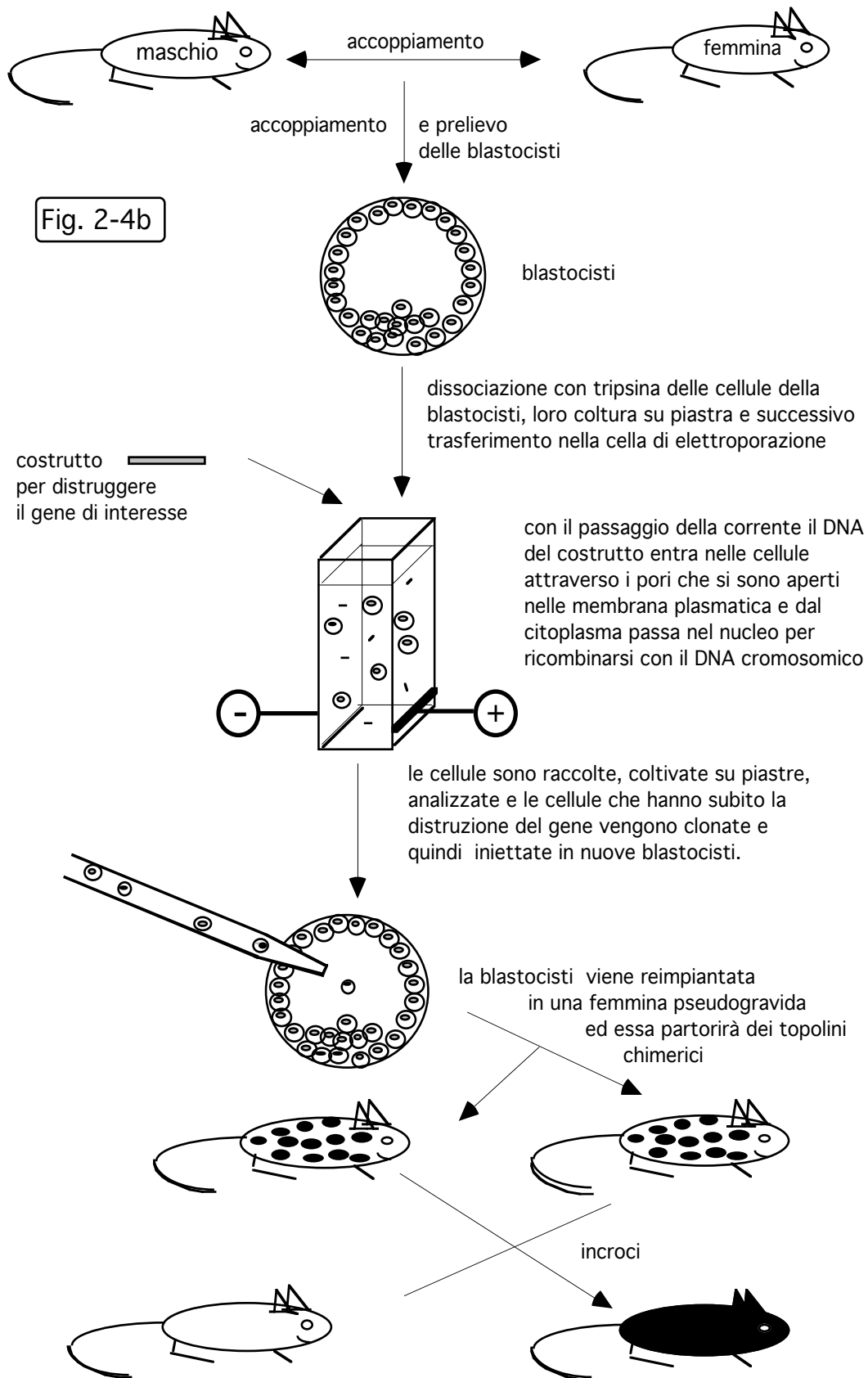



Figura2-4b, didascalia nella pagina precedente

Oligodesossinucleotidi formanti una tripla elica.
(Triplex Forming Oligonucleotides)
Oligodesossinucleotidi a singolo filamento (TFO) si possono legare alla scanalatura maggiore del DNA formando una tripla elica mediante legami ad H, detti legami ad H di Hoogsteen. In natura DNA a tre filamenti è stato osservato nel DNA mitocondriale nell'unica regione non codificante di DNAds (D-loop) a cui si associa un breve filamento di DNA (7S DNA). La sequenza dei TFO è disegnata per associarsi specificamente al DNAds delle regione promotrice del gene di interesse. Il TFO transfettato nelle cellule migra nel nucleo, si associa al DNAds ed inibisce la trascrizione impedendo il legame alle sequenze promotrici dei fattori di trascrizione. Questa tecnologia ha dei limiti: per inibire la trascrizione occorre una alta concentrazione di TFO, i legami a H di Hoogsteen possono formarsi solo quando la sequenza ha tutte le purine su un filamento, e per stabilizzare i TFO occorre modificarli agli estremi 5' e 3' e ciò ne riduce la specificità.

Oligodesossinucleotidi e geni antisenso. In Natura esistono esempi di RNA antisenso che regolano specificamente l'espressione di geni e questa osservazione ha suggerito di sintetizzare "Oligodesossinucleotidi di DNAss antisenso" (ODN) aventi la propria sequenza complementare al mRNA codificante una proteina di interesse. L'ODN viene transfettato nelle cellule di interesse, si ibrida al mRNA ad esso complementare ed inibisce specificamente la sintesi della proteina a livello della sua traduzione.
Gli oligodesossinucleotidi antisenso sono facilmente sintetizzati *in vitro*, e quando ibridati al mRNA ne favoriscono la distruzione per azione di nucleasi che selettivamente distruggono l'RNA degli ibridi DNA-RNA. Anche il DNA antisenso è degradato dalle cellule, la sua vita può essere allungata con modificazioni chimiche al 5' ed al 3' (atomi di solfo al posto di quelli fosforo). Gli ODN possono migrare nel nucleo ma la loro sequenza è disegnata in modo da avere scarsa possibilità di formare triple eliche con il DNAds cromosomico.
Sono stati costruiti plasmidi contenenti geni antisenso che transfettati in cellule sintetizzano continuamente mRNA antisenso che essendo complementari al mRNA del gene di interesse si ibrideranno ad esso bloccandone la traduzione perché i ribosomi non possono accedere al mRNA o perché l'RNAds è degradato dagli enzimi ribonucleasi della cellula.
Gli ODN ed i geni antisenso permettono di avere informazioni sulla funzione di un gene osservando le alterazioni delle funzioni cellulari conseguenti l'inibizione dell'espressione del gene a livello della traduzione. Gli effetti sulla fisiologia cellulare sono simili o identici a quelli della distruzione del gene che è operabile solo su animali, mentre ODN e geni antisenso possono essere utilizzati per scopi terapeutici anche sull'uomo.
ODN sono stati usati anche in terapia, detta "terapia antisenso" ed è una terapia genica con azione inibitoria, cioè che tende a rimuovere gli "eccessi di attività molecolare" provocati da mutazioni (es. di oncogeni). L'azione terapeutica può essere anche allele specifica quando un allele ha subito una

cospicua alterazione di sequenza come in alcune patologie dominanti per cui si può progettare un oligonucleotide che si leghi solo al mRNA mutato.
L'azione sulla fisiologia cellulare può provocare anche una attivazione di qualche funzione cellulare se il DNA antisenso blocca la sintesi di una proteina con funzioni inibitorie. Si ritiene che gli ODN abbiano un'alta potenzialità terapeutica e per alcuni di essi sono in corso prove cliniche su pazienti.

RNA di interferenza.
Lo RNA di interferenza (**RNAi**) è RNA a doppio filamento (RNAds) che è capace di sopprimere l'espressione di un gene che abbia la sequenza identica alla sua.
Questa capacità è detta "interferenza da RNA" o "riduzione al silenzio di un gene a livello post-trascrizionale" (**PTGS** = post-transcriptional gene silencing).
Lo RNAi non agisce come tale ma modificato da una serie di reazioni covalenti e di associazione.
Quando un RNAds entra in una cellula è attaccato da un enzima chiamato Dicer (ribonucleasi specifica per RNAds) che lo taglia in frammenti di 19b (corrispondono a due giri di doppia elica) con agli estremi 3' altre due basi (in totale i singoli filamenti hanno 21b ma sono appaiati solo per 19b). Lo RNAds si associa a un complesso di proteine che dissociano i due filamenti e favoriscono l'associazione al mRNA endogeno e se l'appaiamento è esatto l'mRNA è distrutto da un ribonucleasi.
Questi frammenti di RNAds di 19b +2b sono indicati con l'acronimo **siRNA** (short interfering RNA) ed il complesso siRNA-proteine che distrugge l'mRNA con l'acronimo **RISC** (RNA-induced silencing complex).
Alcuni autori considerano che il meccanismo molecolare con il quale viene inibita specificamente l'espressione di un gene a livello della traduzione dai RNAi sia lo stesso (stesso complesso di proteine) di quello con il quale operano i geni antisenso (paragrafo precedente).
I siRNA migrano nel nucleo ed interferiscono anche con la trascrizione dei geni degli eucarioti con un meccanismo non ancora completamente definito indicato con la sigla (**TGS** = transcription gene silencing). Si ipotizza che i siRNA possano agire inibendo la trascrizione associandosi al DNA del gene ad essi complementare o alla catena nascente del mRNA.
Questo sofisticato meccanismo di risposta cellulare è universale (presente in piante, funghi, animali) pertanto deve avere una funzione molto importante.
Sono state fatte alcune ipotesi.
Molti virus delle piante e degli animali hanno il genoma di RNAds o di RNAss. Lo RNAss virale una volta introdotto nella cellula è rapidamente convertito in RNAds. Il meccanismo cellulare di formazione degli RNAi avrebbe la funzione di distruggere gli mRNA virali per inibire la sintesi delle proteine virali e quindi la replicazione del virus. Gli RNAi avrebbero anche la funzione di distruggere gli RNA senza funzione trascritti da trasposoni, che talvolta hanno regioni a doppio filamento. Alcuni esperimenti suggeriscono che gli RNAì possano avere un ruolo nel controllo dell'espressione dei geni. L'inibizione della trascrizione da parte di RNAss, in questo caso di grossa dimensione, è responsabile

dell'inattivazione dell'espressione della quasi totalità di geni di uno dei due cromosomi X dei nuclei femminili (appendice D).
Gli RNAi sono utilizzati per inibire l'espressione di singoli geni. RNAi transfettati in ovociti e cellule hanno permesso di inibire specificamente l'espressione di singoli geni. La sequenza del RNAi è disegnata in modo da interagire specificamente con il gene di interesse e non con altri geni. L'inibizione dell'espressione genica è temporanea, tuttavia essa dura sufficientemente a lungo da permettere di osservare il fenotipo alterato ed avere indicazioni sulla funzione del gene. La durata dell'inibizione può essere aumentata fino a almeno due mesi introducendo un vettore che sintetizza continuamente siRNA. Con questo procedimento è stata costruita una pianta di caffè transgenica in cui l'espressione di un gene necessario per la sintesi della caffeina, era inibita da siRNA prodotto da un transgene. Lo scopo è quello di produrre caffè decaffeinato evitando di dover estrarre la caffeina dai chicchi di caffè. siRNA specifici per il gene dell'apoliproteina B (proteina sintetizzata dal fegato ed intestino, trasportatrice di lipidi nel sangue) iniettati endovena nel topo causano la riduzione della sintesi della proteina. Sono già iniziati gli studi per utilizzare RNAi anche nella terapia umana.

Ribozimi. I ribozimi sono molecole di RNA dotate di attività catalitica che può agire sulla loro stessa molecola (autocatalitica) e/o su altre molecole di RNA (transcatalitica). In natura esistono varie molecole di RNA che hanno questa attività catalitica e svolgono un ruolo nel metabolismo degli RNA. I ribozimi hanno una regione che per complementarità di basi lega specificamente una regione della molecola di RNA bersaglio ed un’altra regione con attività catalitica di idrolisi del legame fosfodiestere.
Ribozimi per uso sperimentale o terapeutico sono stati modificati nella parte legante il substrato in modo che possano associare specificamente molecole di mRNA di interesse. Il ribozima modificato è transfettato nelle cellule di interesse dove assocerà l'mRNA di interesse, ne catalizzerà il taglio e così lo esporrà alla distruzione da parte delle RNAasi cellulari che riconoscono gli estremi tagliati del RNA.
L'azione di un ribozima blocca la traduzione di uno specifico mRNA determinandone la degradazione e dalle alterazioni fisiologiche che ne conseguono si hanno informazioni sulla funzione del gene che codifica l’mRNA. Da tempo sono in corso prove cliniche su pazienti trattati con ribozimi che attaccano mRNA di oncogeni.

Analisi dell'espressione di singoli geni. L'analisi dell'espressione di singoli geni nei vari tessuti dell'organismo può essere fatta mediante analisi Northern o Western (capitolo 1).
L'analisi Western, che valuta la concentrazione di una data proteina, è considerata più valida per valutare l'espressione del relativo gene rispetto all'analisi Northern, che valuta la concentrazione del mRNA. Questo perché nelle cellule le concentrazioni relative di molti mRNA (% della concentrazione totale degli mRNA) non corrisponde alla concentrazione relativa della proteina da essi

codificata (% della concentrazione totale delle proteine). In alcuni casi si è osservato che la presenza nelle cellule di un dato mRNA era associata all'assenza completa della proteina codificata. Si è così appurato che esistono geni normali che vengono trascritti ma il loro messaggero non viene tradotto. Questa discrepanza tra la concentrazione del RNA e quella della proteina codificata può risultare dai meccanismi di controllo della traduzione: stabilità del mRNA e della proteina, affinità del mRNA per i fattori del complesso d'inizio della sintesi proteica (appendice B). Anche le analisi che indicano la presenza nella cellula di una data proteina devono essere considerate con prudenza, perché con la tecnologia della distruzione del gene si è osservato che alcuni tessuti, in cui è espressa e presente una data proteina, non subivano alterazioni con la perdita totale della proteina stesse. Ciò è stato interpretato assumendo che alcune proteine presenti nelle cellule svolgano funzioni che possono essere compensate da altre proteine. Secondo un'altra ipotesi, si ritiene che la sintesi di una proteina non utile alla cellula sia tollerata quando la proteina, anche se abbondante, non ha effetti tossici per la cellula. Si ipotizza che la costruzione di un sistema di regolazione per inibire la sintesi di una proteina richieda un costo energetico superiore alla sintesi di una proteina inutile ed innocua per la normale fisiologia delle cellule di alcuni tessuti.
L'espressione di un gene in tutte le cellule dell'organismo suggerisce che la proteina da esso codificata abbia una attività molecolare indispensabile alla vita della cellula, come gli enzimi che mantengono in vita la cellula (house keeping enzymes) (es. enzimi della glicolisi); mentre l'espressione di un dato gene solo in un dato tipo di cellula suggerisce che la proteina sintetizzata in quel tipo di cellula sia coinvolta in una particolare funzione. Se una proteina è sintetizzata esclusivamente in un dato tipo di cellula endocrina, si ha l'indicazione che la proteina possa essere coinvolta nella sintesi, regolazione della sintesi o escrezione dell'ormone. La sintesi di una proteina specificamente in fase G_0 (quiescenza) o S (sintesi del DNA) del ciclo cellulare suggerisce che essa abbia un ruolo rispettivamente nel mantenimento della quiescenza e nella replicazione del DNA.

Tecnologie per l'analisi simultanea dell'espressione di più geni

I microarray.

Le biotecnologie precedentemente descritte, permettono di studiare un gene per volta, con il procedere del progetto genoma e la disponibilità della sequenza di molti geni è stata sviluppata la tecnologia dei microarray (microschiere), al fine di studiare simultaneamente più geni o tutti i geni umani (figura 2-5).
Un tipo di microarray è costituito da una superficie di vetro su cui sono disposte ordinatamente piccole quantità di DNA. Su una superficie di 1,28x1,28 cm può essere depositato il DNA di oltre 60.000 frammenti diversi ciascuno su una posizione diversa e nota di una griglia virtuale avente una disposizione a rete ortogonale. Pertanto mediante una ibridazione effettuata con una sonda di DNA marcato (fluorescente) su un singolo microarray si potranno vagliare

simultaneamente tutti i geni umani (circa 30.000). Un metodo per disporre le varie soluzioni di DNA opera mediante apparecchiature simili a stampanti a getto di inchiostro che depositano piccolissime quantità (0,4nl) di DNA in soluzione. I centri delle micro-gocce distano 120μm l'uno dall'altro. Le micro-gocce vengono essiccate ed il loro DNA rimane aderente al supporto di vetro costituendo delle micromacchie di DNA (DNA microspot). Esiste anche un metodo che permette la sintesi di oligonucleotidi direttamente sul supporto solido del microarray secondo sequenze prestabilite e diverse per ogni punto della griglia. Con questo metodo sono stati costruiti microarray (1,28x1,28cm) contenenti fino a 300.000 oligonucleotidi diversi ancorati su un unico supporto solido. I DNA-microarray costituiti da oligonucleotidi sintetizzati direttamente sul supporto solido sono chiamati DNA-chips per il complesso procedimento di costruzione simile a quello dei chips di silicio dei circuiti del computer.
Sono stati costruiti anche microarray di proteine e l'identificazione dei vari polipeptidi è fatta mediante anticorpi.
Le maggiori applicazioni dei microarray sono: l'analisi dei livelli di espressione dei geni e la ricerca delle variazioni di sequenza nel DNA delle varianti dei geni normali e patologiche.

Analisi dei livelli di espressione di più geni con i microarray.
Un metodo per operare l'analisi dei livelli di espressione di più geni consiste nel preparare i microarray ponendo ordinatamente su supporto solido oligonucleotidi sintetici diversi (DNAss). In ogni posizione della griglia ideale ortogonale sono poste molte copie di un dato oligonucleotide e la posizione è registrata associata con la sigla dell'oligonucleotide. Ogni oligonucleotide è complementare ad un singolo cDNA sintetizzato da un mRNA umano di un dato tessuto oppure a più tessuti umani. Pertanto un DNA sonda che ibridi con un dato oligonucleotide può essere identificato dalla posizione dell'oligonucleotide con il quale si è ibridato. Se la soluzione di ibridazione contiene più cDNA, anche tutti quelli di un tessuto, ad ibridazione avvenuta, si avrà l'indicazione di quanti e quali cDNA erano presenti nella soluzione.
Al fine di avere dati quantitativi, la quantità di oligonucleotide presente in ogni singola posizione è volutamente molto alta (10^6-10^9 copie) affinchè essa sia difficilmente saturata (completamente ibridata) dalle quantità di cDNA presenti nelle soluzioni di ibridazione. Se le copie degli oligonucleotidi nelle singole posizioni fossero poche l'ibridazione avverrebbe egualmente ma sarebbe solo qualitativa (presenza o assenza di ibrido) perché cDNA sonda presenti in quantità diverse si ibriderebbero completamente con gli oligonucleotidi del microarray (saturazione) e non in proporzione alla quantità degli mRNA sonda. Le quantità degli ibridi nelle diverse posizioni del microarray risulterebbero tutte uguali.
Si opera l'ibridazione stendendo sul microarray una soluzione contenente sonde costituite da cDNA sintetizzati da tutti gli mRNA di un dato organo e poi legati covalentemente con una molecola fluorescente. Sono stati usati anche mRNA legati ad una molecola fluorescente. Ciascuna posizione del microarray è

microarray
ibridato

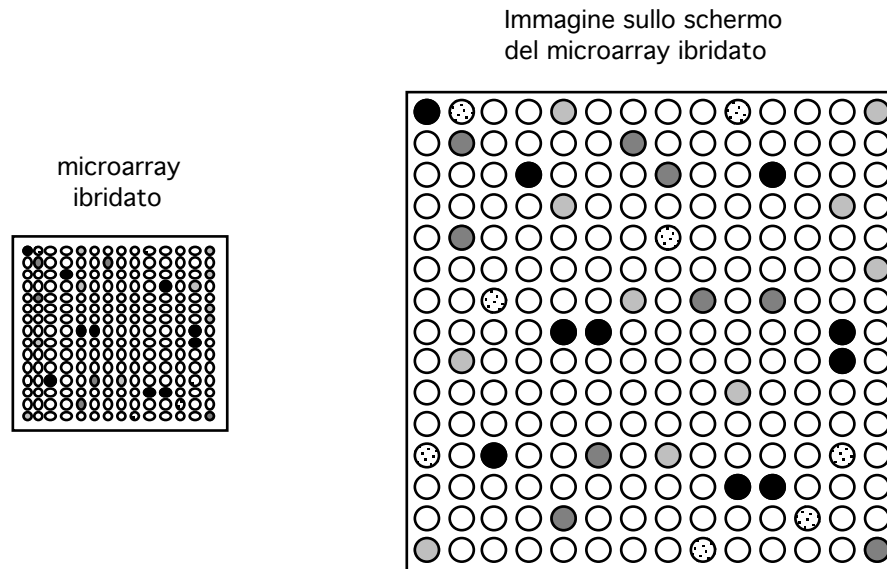



Figura 2-5. Un tipo di microarray (microschiera) è costituito da un supporto di vetro, trattato chimicamente, su cui sono disposti ordinatamente come su una scacchiera delle microquantità di DNA. In ogni posizione è presente lo stesso numero di copie di un oligodesossinucleotide capace di ibridarsi specificamente ad un dato mRNA (o cDNA) di un dato tessuto (es. fegato umano sano). Sul microarray viene stesa una soluzione contenente i cDNA sintetizzati da mRNA totali estratti dallo stesso organo (fegato umano sano) e resi fluorescenti mediante un fluorocromo legato covalentemente ad ogni molecola di cDNA. Microarray e soluzione sono posti nelle condizioni ottimali di temperatura affinché avvengano le ibridazioni tra gli oligodesossinucleotidi del microarray e i cDNA in soluzione. Scaduto il tempo necessario per l'ibridazione, il microarray viene lavato per eliminare la soluzione con i cDNA sonda che non si sono ibridati, e sottoposto a scansione ad alta risoluzione. Un microfascio di una radiazione laser avente la lunghezza d'onda, idonea a sollecitare la fluorescenza del fluorocromo legato ai cDNA, colpisce ogni singola micromacchia ibridata o non ibridata del microarray. Le emissioni fluorescenti sono raccolte e convertite in una immagine amplificata del microarray ibridato su uno schermo (monitor) di un apparecchio di controllo. I gradi di intensità della fluorescenza di ogni ibrido sono indicati con intensità diverse di uno stesso colore o con colori diversi. L'intensità della fluorescenza, che può essere convertita anche in valori numerici, indica la quantità dell'ibridato che si è formato e quindi la quantità del cDNA fluorescente presente nella soluzione di ibridazione, dato che la quantità del cDNA ancorato al vetro del microarray è talmente alta in tutte le micromacchie da non venir saturata neanche dai cDNA sonda più abbondanti. La quantità del cDNA corrisponde alla concentrazione cellulare della specie molecolare del mRNA da cui è stato sintetizzato ogni cDNA. Quindi il grado di fluorescenza emessa da ogni singola micromacchia ibridata indica l'intensità dell'espressione di ogni singolo gene del tessuto analizzato (es. fegato umano normale). Utilizzando una soluzione di ibridazione contenente i cDNA fluorescenti di un carcinoma epatico e lo stesso microarray si ha l'indicazione dei geni espressi nelle cellule cancerose. Sottraendo dai geni espressi nel carcinoma epatico quelli espressi nel fegato normale si ottengono i geni espressi specificamente nel carcinoma epatico e tra essi possono essere ricercati i geni responsabili della patologia.

esplorata con una macchina dotata di un microraggio laser; dove si sono formati ibridi la molecola fluorescente viene eccitata dal raggio laser e la sua radiazione è registrata dalla stessa macchina che la elaborata e la trasmette amplificata su un monitor (figura 2-5).
La posizione sul microarray dell'ibridato identifica il cDNA sonda presente nella soluzione di ibridazione e quindi del relativo gene.
L'intensità della fluorescenza, che è in relazione diretta alla quantità di ibridi formati in ogni singola posizione, rivela il grado di concentrazione del cDNA sonda ibridato, quindi del mRNA da cui è stato sintetizzato e quindi il grado di espressione del relativo gene.
Le condizioni di ibridazione per ogni microarray, ed in particolare la T, non possono essere ottimali per la formazione di tutti i complessi oligodesossinucleotidi-cDNA dato che i vari ibridi hanno sequenze diverse e quindi hanno anche valori diversi di Tm. Può risultare che quantità diverse di fluorescenza (diverse quantità di ibrido) in due diverse posizioni del microarray, risultino per quantità uguali di due mRNA aventi Tm diverse (es. un mRNA ha la Tm uguale alla T di ibridazione mentre l'altro mRNA ha una Tm più bassa ed ibriderà meno efficientemente). Pertanto vengono operate delle correzioni sui dati quantitativi.
Il metodo è comunque eccezionale perché permette di analizzare simultaneamente qualitativamente e con buona approssimazione anche quantitativamente l'espressione dei geni in un dato tessuto.

I microarray sono utilizzati per analizzare qualitativamente e quantitativamente l'espressione di geni :
- di cellule/tessuti/organi sottoposti a particolari condizioni fisiologiche (digiuno, alimentazione, sforzo fisico, azione di ormoni)
- durante lo sviluppo embrionale, l'organogenesi, il differenziamento cellulare.
- durante il ciclo cellulare.
- codificanti gli enzimi di una via metabolica.
- di un dato tessuto affetto da una data patologia
- di un tessuto che è stato esposto a farmaci, anestetici, allergeni o sostanze tossiche.

<u>Analisi delle variazioni di singole basi nella sequenza nel DNA.</u>
Per questo tipo di analisi il microarray è costituito da oligonucleotidi di 20b. Il primo nucleotide è complementare alle basi 1-20 del filamento senso del DNA del gene di interesse, il secondo oligonucleotide è complementare alle basi 2-21 e così avanti fino a coprire tutta la sequenza del gene. Ognuno di questi oligonucleotidi esiste in 4 varianti: una ha la sequenza identica a quella del gene di interesse ed ognuna delle altre tre ha, rispetto alla sequenza del gene, una delle tre possibili basi sostituita nella parte centrale (posizione 10). In questo modo sono prese in considerazione tutte le 4 possibili varianti di ciascuna base

GATTCAGCAGTCGCATCTCAGACACCAACCACTATGCTGTCAGCAGTTGCCCGGGGCTACC
GATTCAGCA**G**TCGCATCTCA
GATTCAGCA**A**TCGCATCTCA
GATTCAGCA**T**TCGCATCTCA
GATTCAGCA**C**TCGCATCTCA
 ATTCAGCAG**T**CGCATCTCAGA
 ATTCAGCAG**C**CGCATCTCAGA
 ATTCAGCAG**G**CGCATCTCAGA
 ATTCAGCAG**A**CGCATCTCAGA

della sequenza del gene. Questi oligonucleotidi sono oligonucleotidi allele specifici (ASO) perché permettono di rivelare la mutazione anche quando il DNA sonda include il DNA di ambedue gli alleli del gene di interesse. In genere il DNA del gene di interesse è diploide perché è estratto o amplificato da cellule diploidi e la possibile mutazione potrà essere omozigote o eterozigote. La sonda fluorescente è costruita amplificando mediante PCR la regione del gene di interesse (la regione promotrice o un suo esone,) utilizzando primer legati covalentemente ad una sostanza fluorescente. Il DNA sonda non mutato si ibriderà solamente e con tutti gli oligonucleotidi perfettamente complementari ad esso (più posizioni nel microrray in relazione della lunghezza della sequenza sonda). Il DNA mutato (sonda) si ibriderà solo con un oligonucleotide se in esso è presente una sola mutazione puntiforme. Essendo nota la posizione degli oligonucleotidi nel microarray, dalla posizione degli ibridi che si sono formati si deduce l'assenza o la presenza di mutazioni, ed in presenza di mutazione, la base mutata e la sua posizione nella sequenza. Questa tecnica è accurata per mutazioni omozigoti, mentre può perderne alcune eterozigoti e ha difficoltà ad individuare mutazioni per inserzione di una base se il microarray non include oligonucleotidi contenenti l'inserzione.
Dato che le mutazioni di un gene possono differire per la base mutata o per la posizione fisica nella sequenza del DNA (come nel gene BRCA1), i microarray rendono più veloce la ricerca delle mutazioni dei geni, rispetto all'esecuzione dell'analisi SSCP o alla determinazione della sequenza di tutti gli esoni di un gene. Tuttavia l'analisi con i microarray può essere utilizzata solo se in precedenza è già stata stabilita la sequenza non mutata del gene di interesse, Inoltre l'analisi della sequenza rimane l'analisi più sicura, sempre utilizzata per confermare i dati.
Con l'analisi su microarray è stato possibile individuare il 90% delle mutazioni dell'oncogene BRCA1.

Alcune considerazioni sulle tecnologie del DNA per ricercare l'attività molecolare e la funzione fisiologica di una proteina

La transfezione di un gene umano in cellule umane in cultura o nell'animale transgenico informa sulla funzione fisiologica del gene determinando un eccesso di proteina codificata che provoca nelle cellule un "incremento della funzione".

La distruzione del gene in tutte le cellule di un animale, i ribozimi specifici per un dato mRNA, lo RNAi ed il DNA antisenso introdotti in cellule in cultura informano sulla funzione fisiologica del gene, determinando la carenza o l'assenza della proteina codificata dal gene di interesse e provocando la "perdita della funzione".
Il gene reporter permette di valutare il momento fisiologico (stadio embrionale, fase del ciclo cellulare, ecc.) ed il luogo fisiologico (organi, ghiandole, nuclei di cellule nervose, ecc.) in cui è sintetizzata la proteina codificata dal gene di interesse.
Queste tecnologie utilizzando la molecola di un gene permettono di ottenere informazioni sulla funzione che la proteina da esso codificata ha nella cellula e nell'organismo. Tecnologie simili non esistono o sono molto più laboriose se si vuole operare con molecole proteiche purificate. Anche la purificazione di una proteina da cellule può essere molto difficoltosa, specialmente se si ha scarsa disponibilità di cellule/tessuti da cui estrarla e questo è frequente per i tessuti umani. Attualmente si preferisce purificare (clonare) il gene, inserire il gene in vettori di espressione e sintetizzare *in vitro* la proteina, soprattutto quando della proteina è conosciuta solo l'attività molecolare. Con particolari accorgimenti (gene chimerico) si semplificano molto le procedure di purificazione delle proteine. Ad esempio può essere costruito un cDNA chimerico costituito dal cDNA del gene di interesse legato a DNA esogeno che codifica un dominio proteico che promuove l'escrezione della proteina dalle cellule transfettate. Il dominio esogeno è rimosso per proteolisi, prima che la proteina sia escreta nel mezzo di coltura. La proteina concentrata è poi estratta dal mezzo di coltura. Inoltre le proteine sono più difficili da purificare e manipolare perché in genere sono molto più fragili del DNA. Molte proteine perdono irreversibilmente la loro struttura terziaria (denaturazione) se sono troppo diluite, se sono esposte a temperature anche poco superiori ai 37°C o a pH non troppo vicini alla neutralità o se congelate. Tutto ciò non accade per il DNA che, anche se denaturato ad alte temperature, riacquista poi la sua naturale struttura a doppio filamento.
La stabilità molecolare del DNA e la potenza delle biotecnologie ha portato ad intraprendere e a completare rapidamente il progetto genoma umano (identificazione tutti i geni umani) con la convinzione che l'utilizzazione delle tecnologie del DNA sia la via più rapida per arrivare a conoscere la funzione fisiologica e l'attività molecolare di tutte le proteine umane.
La conoscenza della sequenza nucleotidica del gene e di quella aminoacidica della proteina forniscono utili ma indirette indicazioni (non prove) sulla funzione del gene e del suo prodotto, perché dedotte da omologie strutturali di geni e di proteine già noti.
Le tecniche del DNA ricombinante possono dare informazioni decisive per definire la funzione fisiologica che l'attività molecolare di una proteina svolge nelle cellule, nello sviluppo embrionale e nell'organismo adulto. Permettono di determinare la sequenza aminoacidica della proteina codificata dal gene deducendola dalla sequenza del cDNA ed in questo modo identificano la

molecola proteica. Permettono di sintetizzare *in vitro*, mentre raramente forniscono indicazioni sull'attività molecolare delle proteine e quindi anche sul possibile dosaggio di questa attività (Tabella 2-1).
Una proteina ha almeno tre caratteristiche molecolari, tutte dipendenti dalla sequenza aminoacidica, che la identifica: una è l'identità molecolare data dalla sequenza aminoacidica (la sua formula di struttura), la seconda è l'attività molecolare che è stabilita quando la proteina assume spontaneamente la sua naturale conformazione terziaria/quaternaria, la terza è la funzione fisiologica cellulare che la proteina svolge nella cellula operando con la sua attività molecolare. Una proteina con la stessa attività molecolare, può svolgere funzioni diverse in cellule di tipo diverso. Esistono anche proteine che hanno attività molecolari diverse e quindi anche funzioni fisiologiche diverse in cellule diverse o in compartimenti diversi di una stessa cellula. (appendice A).
La sequenza aminoacidica identifica la molecola proteica.
L'attività molecolare identifica la natura biologica di una proteina ed in genere dà il nome alla proteina. L'enzima "glucoso-6P-deidrogenasi" è così chiamato perché catalizza la reazione di deidrogenazione del glucosio-6P.
Il dosaggio specifico dell'attività molecolare è l'unica tecnologia che permette di identificare l'attività biologica della proteina. E quindi permette di individuare le basi molecolari responsabili della funzione fisiologica della proteina.
Per mostrare le differenze esistenti tra ricerca dell'attività molecolare di una proteina e la sua funzione nella cellula utilizzo alcuni esempi. Clonato un nuovo gene umano, dedotta la sequenza aminoacidica della proteina codificata, se la proteina non appartiene ad una famiglia di proteine note o non ha domini di nota attività molecolare non si hanno indicazioni sull'attività molecolare della proteina, cioè la sequenza aminoacidica non dà alcuna informazione sull'attività molecolare della proteina. Mentre se la proteina nella sua sequenza ha, ad esempio, i tratti che caratterizzano la famiglia delle deidrogenasi (dominio di legame del NAD/NADP e gruppi responsabili della catalisi), si ha l'indicazione (non la prova) che il gene codifichi per un enzima con attività catalitica deidrogenasica ma non si hanno indicazioni sul possibile substrato e cofattore specifico (NAD o NADP) dell'enzima.

Tabella 2-1.

ANALISI E SINTESI DI	PROTEINE TECNOLOGIE BIOCHIMICHE	ACIDI NUCLEICI BIOTECNOLOGIE
Taglio specifico	enzimi proteasi (pochi punti di taglio specifico)	enzimi di restrizione (molti punti di taglio specifico)
Unione specifica di frammenti	nulla	ligasi
Possibilità di ricombinazione *in vitro*	nulla	altissima
Rivelazione	*dosaggio dell'attività biologica, anticorpi,	ibridazione, PCR microarray saggio di complementazione
Analisi parziale della struttura molecolare	fingerprinting	mappa di restrizione
Analisi totale	sequenziatore: circa 20 amino acidi per volta (laboriosa)	sequenziatore: circa 400 basi per volta (semplice e veloce)
Purificazione	Procedura complessa data la diversità molecolare ed instabilità delle proteine	clonazione
Sintesi *in vitro* senza stampo	sintetizzatore: peptidi di circa 20 amino acidi	sintetizzatore: oligonucleotidi di circa 20 basi
Sintesi *in vitro* con stampo	traduzione di mRNA in sistemi privi di cellule; transfezione del cDNA e sintesi della proteina di interesse. La proteina sintetica può essere resa chimerica con un dominio proteico che semplifica la sua purificazione.	subclonazione e PCR
Mutazione sito-specifica	Mediante mutazione del gene o cDNA poi come sopra.	Sintesi *in vitro* di DNA con primer mutati

Tabella 2-1. La tabella mostra la maggiore potenza delle tecnologie del DNA rispetto a quelle biochimiche. Le biotecnologie permettono di manipolare il DNA ed RNA con maggiore versatilità di quanto non possano fare le tecnologie biochimiche con le proteine. Inoltre le biotecnologie utilizzando il cDNA permettono di operare (mutare, sintetizzare e purificare) sulle proteine (indicato dallo spostamento nella colonna delle proteine della doppia riga degli acidi nucleici).
*Le biotecnologie non possono sostituire alcune tecnologie biochimiche: ad esempio il dosaggio dell'attività molecolare delle proteine ed il dosaggio specifico delle proteina che si fa con gli anticorpi.

Per individuare il substrato, il cofattore ed i possibili effettori allosterici occorrono prove molecolari. Occorre purificare la proteina ed esporla a possibili molecole di substrato, che possono essere centinaia, e di più tipi di cofattore e verificare se e come i substrati sono modificati nella reazione catalizzata dalla proteina. Individuati substrato e cofattore si hanno le conoscenze sufficienti per sviluppare un metodo di dosaggio specifico dell'attività catalitica della proteina per cui si possono poi ricercare le sue molecole regolatrici che ne modificano (aumentano o diminuiscono) l'attività catalitica. Egualmente occorrono prove biochimiche e cellulari per dimostrare che le sequenze segnale presenti nella sequenza aminoacidica della proteina svolgono effettivamente la loro attività molecolare. Talvolta una sequenza segnale sebbene presente nella sequenza aminoacidica di una proteina si trova all'interno della molecola proteica e quindi non è attiva alle analisi molecolari. Lo studio delle proteine è tappezzato di trabocchetti.
Le tecniche del DNA sono molto utili anche per migliorare la conoscenza della funzione di geni clonati mediante la clonazione funzionale (capitolo 4), cioè di geni la cui clonazione è stata possibile perché era già nota la sequenza o l'attività molecolare della proteina da essi codificata. Un esempio: il recettore dei glucorticoidi era una proteina già ben nota nella funzione ed attività e nella localizzazione tissulare (fegato ed in minore quantità nel polmone). Con esperimenti sui topolini, la distruzione del gene del recettore dei glucocorticoidi ha permesso di verificare che il recettore svolgeva una funzione nell'organogenesi del polmone ma non in quella del fegato.

<---------------------------->

Piccolo Glossario

Genoma, l'insieme di tutti i geni di un organismo (unicellulare o pluricellulare) o di un virus. Un' altra definizione: l'insieme delle diverse molecole di DNA dei cromosomi di una specie (nell'uomo le specie molecolari di DNA sono 25: 22 autosomi, 2 cromosomi sessuali ed 1 cromosoma mitocondriale).
Genomica è la disciplina che studia il genoma nel suo complesso.
Rnoma (pronuncia: erreenneoma) anche detto **trascrittoma** è l'insieme di tutti gli mRNA espressi da un organismo, o se indicato, di una singola cellula.
Rnomica è la disciplina che studia l'Rnoma nel suo complesso.
Proteoma è l'insieme di tutte le proteine di un organismo o, se indicato, di una singola cellula.
Proteomica è la disciplina che studia il proteoma.
Genomica funzionale è lo studio su larga scala dell'espressione di tutti i geni di una specie.
Genomica comparata analizza i genomi di specie diverse mediante il confronto dell'intera sequenza dei loro DNA e delle sequenze delle proteine codificate da tutti i geni del genoma.

Poca osservazione e molto ragionamento conducono all'errore.
Molta osservazione e poco ragionamento conducono alla verità.
Alexis Carrel, premio Nobel per la Medicina

Capitolo 3

Tecnologie per definire la posizione cromosomica e subcromosomica dei geni

L'analisi molecolare dei geni umani ha percorso due strade: la definizione della loro intima natura molecolare, che si è conclusa con la determinazione della sequenza dei nucleotidi del DNA dei geni, e la definizione della posizione fisica dei geni nella sequenza nucleotidica delle molecole di DNA dei cromosomi.

I geni hanno almeno tre caratteristiche che li identificano: l'identità molecolare data dalla sequenza nucleotidica, l'informazione genetica che è contenuta nella sequenza dei nucleotidi del DNA dei geni (che si manifesta mediante le attività molecolari delle diverse sequenze che li costituiscono e delle proteine che codificano) e una caratteristica indipendente dalla sequenza dei geni che è la posizione nella sequenza della molecola di DNA del cromosoma sul quale è localizzato il gene: posizione subcromosomica (locus). La conoscenza del locus dei geni è importante perché i geni, differentemente dalle proteine, non sono molecole indipendenti ma allineate e legate in sequenza a formare il DNA dei cromosomi e la loro trasmissione da un organismo all'altro deve seguire regole genetiche. La conoscenza della posizione subcromosomica dei geni permette di seguire la trasmissione dei caratteri genetici normali e patologici da un individuo ai suoi discendenti. La conoscenza della sequenza di un gene permette di esplorare la sua funzione negli organi e nell'organismo.

La definizione della posizione cromosomica e subcromosomica dei geni, che è detta mappatura dei geni, è iniziata con l'analisi del fenotipo mediante la costruzione degli alberi genealogici, delle mappe citogenetiche e di ricombinazione e, con il progetto genoma umano, è continuata con l'analisi sistematica delle sequenze del DNA genomico per concludersi nel 2003 con la costruzione della mappa fisica di tutti i cromosomi del genoma umano. Di seguito vengono descritte brevemente le varie metodologie per la definizione della posizione cromosomica e subcromosomica dei geni.

<u>Costruzione degli alberi genealogici familiari con l'indicazione dei membri portatori di un allele normale o patologico</u>

L'esistenza di un dato gene può essere dedotta confrontando le diversità del suo fenotipo, cioè delle caratteristiche della molecola, dell'attività molecolare o della funzione cellulare della proteina da esso codificata in individui geneticamente diversi tra loro (es. portatori di gruppi sanguigni diversi) oppure confrontando individui sani con individui malati, portatori di una mutazione che altera una funzione fisiologica (figure 3-1 e 3-2).

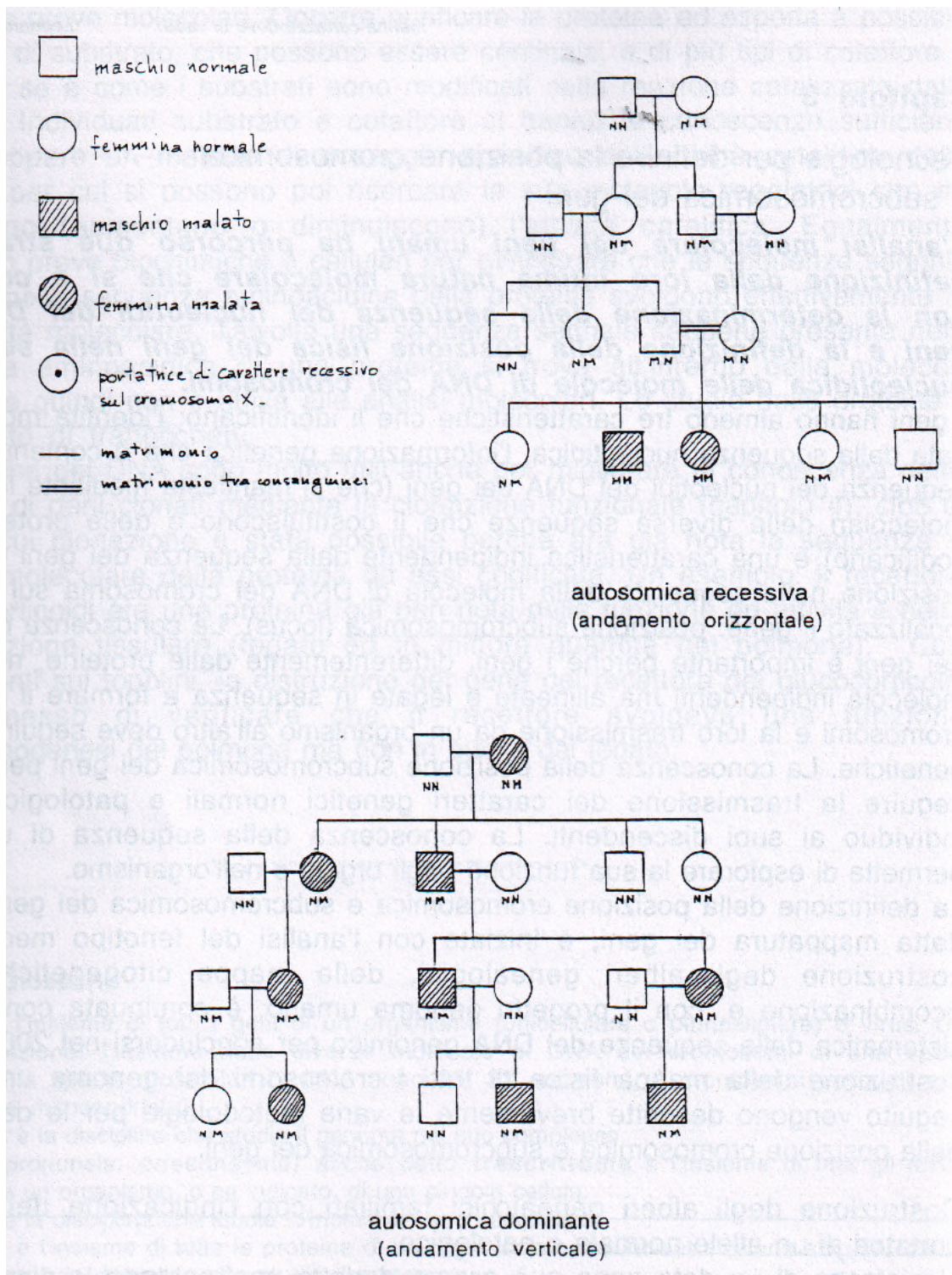



Figura 3-1. Alberi genealogici di malattie autosomiche. N, allele normale; M, allele mutato (responsabile della malattia) (ridisegnato e modificato da Connor J.M. and Fergusson-Smith M. A. (1991) Essential Medical Genetics, 3rd ed., Blackwell, London).

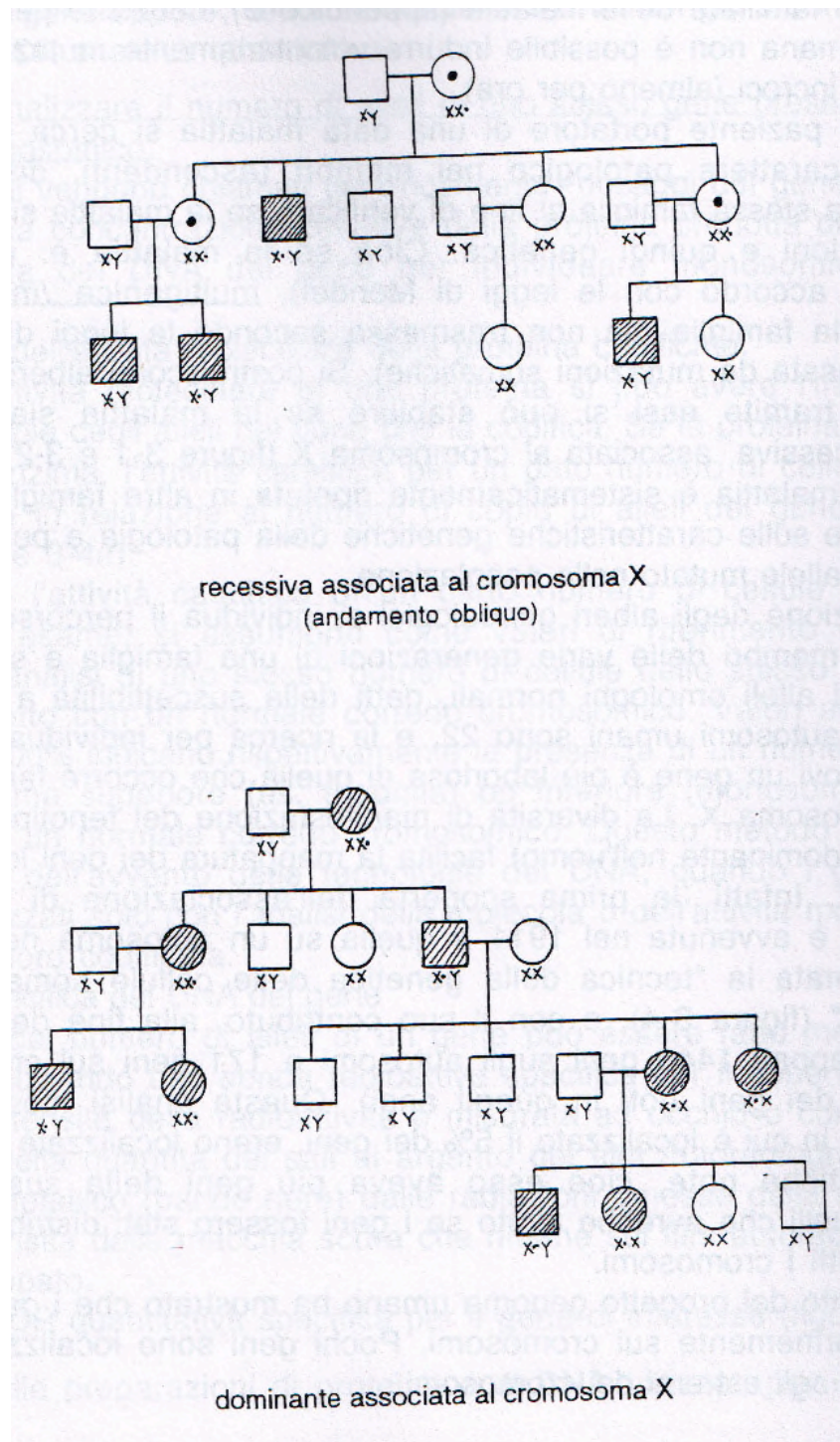



Figura 3-2. Alberi genealogici di malattie associate al cromosoma X. Per altri dati vedere figura 3-1 e testo (ridisegnato e modificato da Connor J.M. and Fergusson-Smith M. A. (1991) Essential Medical Genetics, 3rd ed.,Blackwell, London).

Prima dell'avvento delle tecnologie del DNA, il mezzo storicamente più usato per verificare la presenza di mutanti umani è stata l'osservazione dell'andamento familiare delle malattie (appendice E), dato che per motivi etici nella specie umana non è possibile indurre volontariamente mutazioni, né fare esperimenti di incroci (almeno per ora).
Individuato un paziente portatore di una data malattia si cerca la presenza dello stesso carattere patologico nei membri (ascendenti, discendenti e collaterali) della stessa famiglia al fine di verificare se la malattia sia trasmessa attraverso le generazioni e sia quindi genetica. Cioè se la malattia è: monogenica (trasmessa in accordo con le leggi di Mendel), multigenica /multifattoriale (ricorrente nella famiglia ma non trasmessa secondo le leggi di Mendel) o sporadica (causata da mutazioni somatiche). Si costruiscono alberi genealogici (pedigree) e tramite essi si può stabilire se la malattia sia ereditaria, dominante, recessiva, associata al cromosoma X (figure 3-1 e 3-2). L'indagine per la stessa malattia è sistematicamente ripetuta in altre famiglie al fine di avere conferme sulle caratteristiche genetiche della patologia e per valutare la frequenza dell'allele mutato nella popolazione.
Con la costruzione degli alberi genealogici si individua il percorso degli alleli patologici nei membri delle varie generazioni di una famiglia e si assume la presenza degli alleli omologhi normali, detti della suscettibilità a quella data patologia. Gli autosomi umani sono 22, e la ricerca per individuare su quale autosoma si trovi un gene è più laboriosa di quella che occorre svolgere per i geni posti sul cromosoma X. La diversità di manifestazione del fenotipo (recessivo nella donna e dominante nell'uomo) facilita la mappatura dei geni localizzati sul cromosoma X. Infatti, la prima scoperta dell'associazione di un gene al cromosoma X è avvenuta nel 1911 mentre quella su un autosoma nel 1967. Nel 1967 fu inventata la "tecnica della genetica delle cellule somatiche ibride interspecifiche" (figura 3-4), e con il suo contributo, alla fine degli anni '90, erano stati mappati 1446 geni sugli autosomi e 171 geni sul cromosoma X (circa il 37% dei geni noti in quegli anni). Questa analisi mostrò che nel cromosoma X, in cui è localizzato il 5% dei geni, erano localizzate il 15% delle malattie genetiche note, cioè esso aveva più geni della suscettibilità a patologie di quelli che avrebbe avuto se i geni fossero stati distribuiti in modo uniforme su tutti i cromosomi.
Il completamento del progetto genoma umano ha mostrato che i geni non sono distribuiti uniformemente sui cromosomi. Pochi geni sono localizzati vicino al centromero ed agli estremi dei cromosomi.

Definizione della posizione cromosomica dei geni associati ad aberrazioni cromosomiche responsabili di patologie.

Quando una patologia è associata ad aberrazioni cromosomiche, cioè alterazioni nel numero o nella struttura dei cromosomi, mediante le tecniche di dosaggio del gene si può avere l'indicazione dei geni che mappano sul cromosoma aberrante.

Il dosaggio del gene è la definizione del numero di copie degli alleli di un gene presenti nel genoma ottenuta mediante analisi dell'attività molecolare della proteina codificata dal gene oppure mediante analisi del DNA del gene con le tecnologie Southern o PCR analitica quantitativa.

Metodi per analizzare il numero di alleli di uno stesso gene presenti in un dato corredo cromosomico.

Questi metodi vengono chiamati genericamente "dosaggi del gene" ed operano analizzando la concentrazione cellulare della proteina prodotta dal gene o per analisi diretta del DNA del gene per individuare monosomie, disomie o polisomie.

a. Dosaggio dell'attività molecolare della proteina codificata.

Dosando l'attività molecolare di una proteina si può avere l'indicazione del numero di copie degli alleli del gene che la codifica. Se la proteina codificata dal gene è un enzima, l'attività catalitica per un dato numero di cellule o grammo di tessuto, è in relazione al numero di copie di alleli del gene di interesse (figure 3-3a e 3-4a).

Per valutare l'attività catalitica di un certo numero di cellule prelevate dal soggetto in esame, si assumono come valori di riferimento (100%) quelli ottenuti dall'analisi di uno stesso numero di cellule dello stesso tipo prelevate da un soggetto con un normale corredo cromosomico. Valori al disopra o al disotto del 100% indicano rispettivamente la presenza di un numero di alleli del gene in esame superiore (es. trisomia) od inferiore (monosomia) rispetto a soggetti con un normale corredo cromosomico. Questo metodo è stato molto usato prima dell'avvento delle tecnologie del DNA, quando i geni potevano essere analizzati solo con l'analisi della molecola o dell'attività molecolare della proteina da loro codificata.

b. Analisi specifica del DNA del gene

Il dosaggio del numero di alleli di un gene può essere fatto mediante analisi Southern utilizzando una sonda radioattiva specifica per il gene (figure 3-3b, d e 3-4b). L'intensità della radioattività è misurata ad occhio o con sistemi ottici sulla base della quantità dei sali di argento del film autoradiografico convertiti in argento metallico (bande nere) dalle radiazioni emesse dalla sonda. Cioè si misura l'intensità della macchia scura che rimane sul film autoradiografico dopo averlo sviluppato.

c. mediante PCR quantitativa specifica per il gene di interesse (figura 1-9).

Controllo delle preparazioni di proteine e di DNA utilizzati per il dosaggio del gene.

Il prelievo delle cellule e la successiva manipolazione del DNA possono portare a perdita di proteine e/o DNA di entità diverse nei diversi campioni alterando così i dati quantitativi relativi di proteine e DNA. Pertanto occorre avere sempre dei controlli (detti interni) di ogni preparazione analizzata.

Nell'analisi Southern, come controllo interno, si usa una sonda di DNA che si

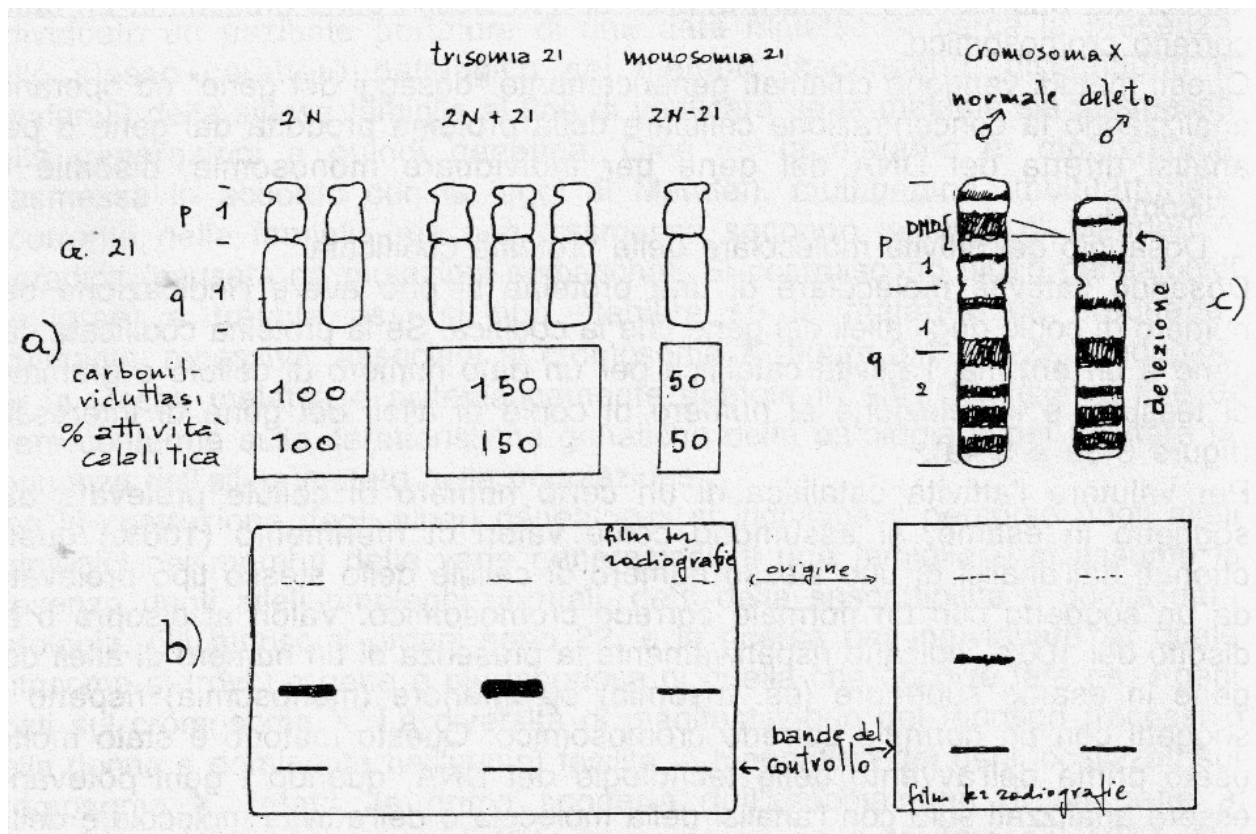



Figura 3-3. Dosaggio del gene mediante: a) dosaggio dell'attività del suo prodotto genico (carbonil-riduttasi, CaR); b) e c) mediante analisi Southern. a) e b) trisomia e monosomia 21; c) delezione del cromosoma X di un individuo maschio portatore di distrofia muscolare di Ducenne (DMD). Per semplicità di disegno, in ambedue le figure in alto, sono indicati solo i cromosomi aberranti (nel numero o nella struttura) responsabili della patologia ed i loro cromosomi omologhi normali. b) in basso nelle autoradiografie si vedono le bande della sonda di controllo che ibrida con un frammento di DNA che mappa su un cromosoma (es. Chr. 1) che non ha subito nessuna aberrazione nei pazienti e nei normali, per cui l'intensità della banda di controllo deve essere uguale nei tre campioni. L'identità dell'intensità delle bande ottenute con le sonde di controllo, indica che i campioni posti sul gel contenevano con buona approssimazione le stesse quantità di DNA totale prelevato dai vari soggetti. Per altri dettagli vedere il testo.

---

ibrida con una regione di DNA, presente nel DNA degli individui normali e dei malati, avente il locus su un cromosoma diverso da quello su cui ha il locus il gene responsabile della patologia. In figura 3-3b, la sonda di controllo è stata scelta in modo che si ibridi con il DNA di un cromosoma diverso dal 21, altrimenti si comporterebbe come la sonda di analisi. In figura 3-3d, la sonda può essere stata scelta in modo che si ibridi sulla parte residua dello stesso cromosoma X o su un altro cromosoma. Un uguale grado di intensità delle bande ottenute con le sonde di controllo, indica che i campioni posti sul gel contenevano le stesse quantità di DNA totale prelevato dai vari soggetti. Se ciò non si fosse verificato, ad esempio se durante la procedure fosse stato perso inavvertitamente il 50% del DNA di un campione, la banda di controllo avrebbe rivelato la perdita perché sarebbe risultata avere metà intensità rispetto alle altre. Utilizzando il valore dell'intensità delle varie bande di

controllo (misurate con un densitometro) e con un semplice calcolo si arriva a normalizzare i valori eliminando le variazioni di intensità anche causate da piccole perdite di DNA nelle diverse preparazioni.

<u>Alcuni esempi della definizione della posizione cromosomica e subcromosomica dei geni associati ad aberrazioni cromosomiche responsabili di patologie.</u>

Consideriamo il caso di un individuo portatore di una trisomia del cromosoma 21 (sindrome di Down). Quando in un dato numero di cellule di quell'individuo viene dosata l'attività catalitica di un enzima codificato da un gene del quale si ricerca la localizzazione cromosomica (es. carbonil-riduttasi, figura 3-3a) si ottiene un valore di attività catalitica del 50% superiore ai valori della stessa attività dosata nelle cellule degli individui aventi un normale corredo cromosomico, mentre negli individui portatori di una monosomia 21 la stessa attività sarà ridotta del 50%. Da ciò si deduce che il gene sia localizzato sul cromosoma 21.

Identici risultati si ottengono se si dosa direttamente il gene di interesse nelle cellule degli stessi individui con le tecniche Southern (figura 3-3b) e PCR analitica quantitativa.

La posizione subcromosomica di un gene può essere ricercata, con l'analisi Southern, ad esempio nel genoma di individui maschi portatori della distrofia muscolare di Ducenne (DMD). La DMD risulta dalla delezione nella parte distale del braccio piccolo (p) del cromosoma X (figura 3-3c). Se utilizzando sonde specifiche all'autoradiografia non risulta alcuna radioattività nel cromosoma X come visto per la DMD (figura 3-3d) e neppure negli altri cromosomi dello stesso individuo portatore della DMD, mentre essa è presente nei cromosomi X normali, si ha l'indicazione che tale gene sia localizzato nella parte deleta del cromosoma X includente il gene della suscettibilità alla DMD.

La storia di questi tipi di mappatura dei geni è sinteticamente questa. Individuati dalle alterazioni del fenotipo (sintomi clinici) gli individui portatori di una stessa patologia genetica appartenenti ad una o più famiglie, i citogenetisti pazientemente hanno costruito alberi genealogici dei membri di ciascuna famiglia (figure 3-1 e 3-2) e pazientemente hanno esaminato e catalogato i loro cromosomi. In alcuni casi, hanno riscontrato che i portatori di una stessa patologia erano anche portatori di una stessa aberrazione cromosomica e da ciò hanno dedotto che l'alterazione cromosomica fosse associata e quindi responsabile della malattia. In questo modo la sindrome di Down è stata associata alla trisomia 21 e la distrofia muscolare di Ducenne alla delezione del braccio p del cromosoma X nei maschi.

Stabilito che i portatori di una data patologia hanno una data aberrazione cromosomica, essi sono stati poi utilizzati per definire la localizzazione cromosomica e subcromosomica di altri geni posti sullo stesso cromosoma responsabile della patologia.

E' importante ricordare che non sempre una stessa malattia è associata alla alterazione cromosomica visibile al microscopio ottico, infatti la mutazione del gene della suscettibilità a tale patologia può essere invisibile al microscopio

ottico perché puntiforme o comunque non sufficientemente grande da alterare la struttura del cromosoma. Ad esempio la distrofia muscolare di Ducenne solo nel 65% dei casi è associata alla delezione del cromosoma X. Nei primi studi, questa non completa associazione tra malattie ed alterazioni cromosomiche visibili al microscopio aveva creato delle indecisioni negli studi diretti a definire la causa genetica di alcune malattie. Tuttavia (come sempre) l'intuito e la costanza di alcuni ricercatori ha permesso di chiarire l'apparente incoerenza esistente tra l'insorgere di una stessa patologia e le cause molecolari apparentemente diverse.

Analisi genetica delle cellule somatiche ibride interspecifiche.
La tecnologia dell'analisi genetica delle cellule somatiche ibride interspecifiche (AGCSI) realizzata nel 1967, permette di individuare il cromosoma e posizione subcromosomica citogenetica nella quale mappa un gene.
Cellule umane e cellule di animali (topo, criceto o scimmia) possono essere fuse a formare cellule ibride mediante l'azione di virus (Sendai) o di agenti chimici (glicole polietilenico). Gli ibridi inizialmente hanno un corredo cromosomico doppio in uno stesso nucleo, le cellule figlie tendono a perdere dal nucleo i cromosomi umani fino a conservarne pochi od anche uno solo. Se si sono utilizzate cellule umane provenienti da un portatore di una aberrazione strutturale di un cromosoma, o da cellule umane precedentemente sottoposte a radiazioni che hanno frammentato dei cromosomi, si possono ottenere ibridi contenenti frazioni di un cromosoma.
Con questa tecnica si ottengono cloni cellulari (cellule identiche perché originate da una singola cellula) contenenti ciascuno differenti cromosomi umani o solo parti di esso (figura 3-4). L'ibrido cellulare uomo-animale, che vive in virtù del patrimonio genetico della cellula animale, permette di avere cellule, con un singolo cromosoma umano o con singole parti di esso, da coltivare quando occorre ed in quantità sufficienti per poter effettuare tutti gli esperimenti che si vuole. Quando non occorrono le cellule vengono congelate e mantenute a T inferiori a -80°C. L'identificazione dei cromosomi umani nelle cellule ibride è possibile osservando la loro morfologia, la dimensione e disposizione delle bande e la colorazione che assumono con specifici coloranti che sono diverse da quelle dei cromosomi di topo. La mappatura del gene, cioè la ricerca del gene nei vari ibridi cellulari contenenti un singolo cromosoma o parte di esso, può essere effettuata mediante dosaggio dell'attività molecolare della proteina codificata dal gene di interesse, mediante analisi Southern con sonde ottenute dal DNA del gene di cui si cerca il locus oppure mediante PCR analitica.
Quando si dosa l'attività molecolare della proteina, le cellule umane sono ibridate con cellule di topo che per mutazione mancano della stessa attività molecolare da dosare, al fine di essere sicuri che l'attività dosata (se presente) sia l'espressione del gene umano, dato che il semplice dosaggio dell'attività molecolare non permette di distinguere se l'attività sia umana o di topo.

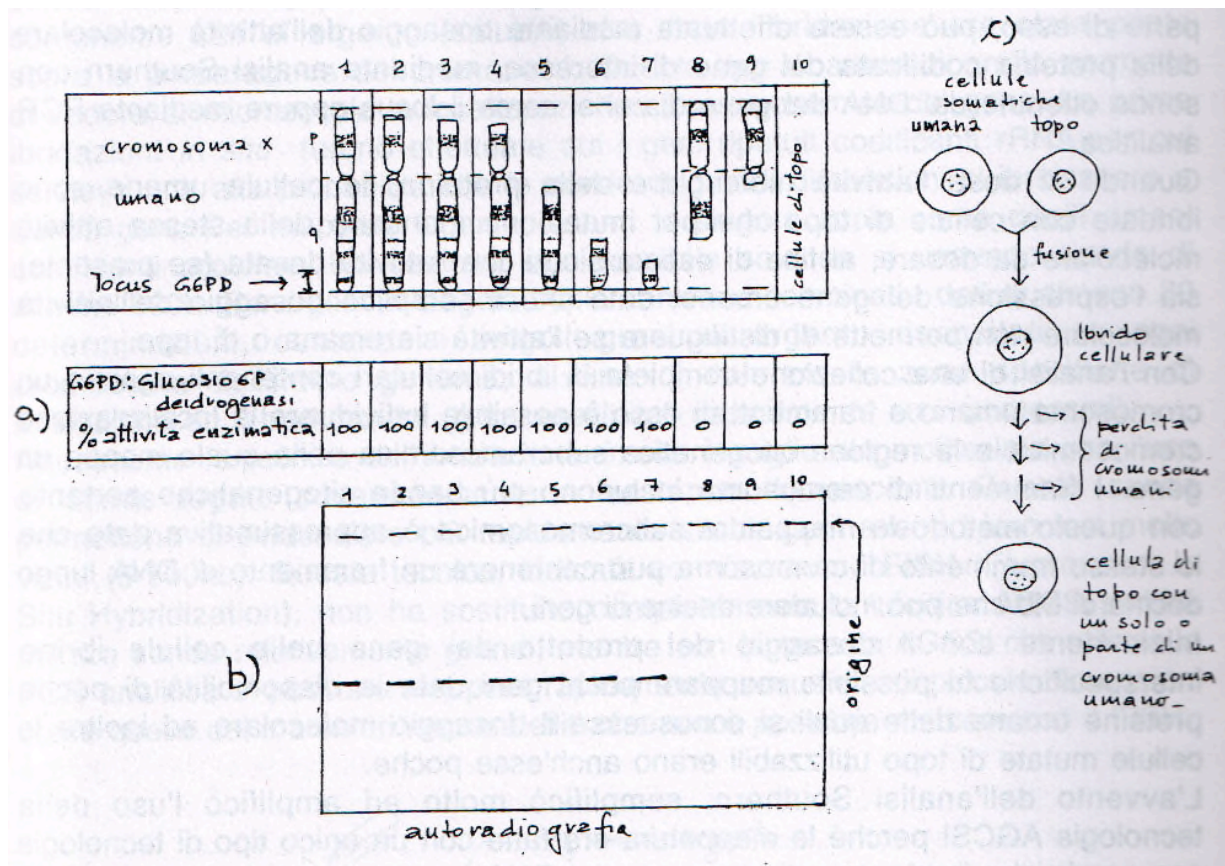



Figura 3-4. Tecnica per l'analisi genetica delle cellule somatiche ibride. a) Mappatura del gene dell'enzima glucosio-6P deidrogenasi (G6PD) sul cromosoma X umano mediante dosaggio dell'attività enzimatica della proteina codificata dal gene e b) mediante analisi Southern. c) costruzione di ibridi cellulari uomo-topo. Nel riquadro in alto sono indicati il cromosoma X e nove differenti parti di esso ottenute mediante radiazione e provenienti da 10 differenti cloni di ibridi cellulari uomo-topo. Sotto in colonna sono riportati i valori del % di attività G6PD e l'autoradiografia di una analisi Southern. Per altri dettagli vedere testo e figura 3-3 (ridisegnato e modificato da Connor J.M. and Fergusson-Smith M. A. (1991) Essential Medical Genetics, 3rd ed.,Blackwell, London).

Con l'analisi di una collezione completa di ibridi cellulari contenenti ognuno un cromosoma umano o frammenti di esso è possibile individuare la localizzazione cromosomica e la regione citogenetica subcromosomica nella quale mappa un gene. I frammenti di cromosoma includono più bande citogenetiche pertanto con questo metodo la mappatura subcromosomica è approssimativa dato che lo stesso frammento di cromosoma può contenere un frammento di DNA lungo decine di Mb che può includere decine di geni.
Inizialmente con il dosaggio del prodotto del gene nelle cellule ibride interspecifiche fu possibile mappare pochi geni data la disponibilità di poche proteine umane delle quali si conoscesse il dosaggio molecolare ed inoltre le cellule mutate di topo utilizzabili erano anch'esse poche.
L'avvento dell'analisi Southern, semplificò molto ed amplificò l'uso della tecnologia AGCSI perché la mappatura era fatta con un unico tipo di tecnologia

con sonde che ibridavano direttamente sul DNA cromosomico e non attraverso dosaggi di proteine, non semplici e diversi per ogni specie di proteina. Successivamente la tecnologia AGCSI è stata ulteriormente semplificata e migliorata dalla possibilità di usare la tecnologia della PCR che ha permesso la mappatura di piccole sequenze su campioni di DNA presente in piccole quantità.
La scoperta delle sequenze ripetute e disperse nel genoma (VNTR, RFLP e SNP) ha permesso la costruzione di una dettagliata mappa citogenetica (vedere dopo) e della mappa di radiazione che è stata molto utilizzata per la costruzione delle mappe fisiche del genoma umano permettendo di ricostruire la sequenza di lunghi tratti di DNA cromosomico (vedere figura 3-12b).

Ibridazione *in* situ.
Per questa tecnica occorre un frammento del gene da mappare (sonda) ed una preparazione di un corredo completo di cromosomi metafasici seccati all'aria, sparsi su un vetrino, in modo da poterli osservare al microscopio. Con una semplice tecnologia (esistono kit commerciali) il DNA costituente la sonda viene reso radioattivo e poi denaturato scaldandolo e raffreddandolo rapidamente. La sonda, messa in contatto con la preparazione di cromosomi (il cui DNA è stato denaturato, privato di RNA ed impoverito di proteine), si ibriderà a sequenze ad essa complementari. La radioattività, emessa dalla sonda ibridata, fa precipitare i grani di argento metallico dall'emulsione contenente sali di argento (solubili) che dopo l'ibridazione era stata posta sopra la preparazione dei cromosomi. Questo tipo di autoradiografia permette di vedere al microscopio la localizzazione subcromosomica del gene. Le prime ibridazioni *in situ* furono effettuate sui i geni ripetuti codificanti rRNA per i quali esisteva una alta probabilità che le sonde di grandi dimensioni si ibridassero a questi geni. Per mappare i geni presenti in singola copia, come quelli umani codificanti proteine, l'ibridazione era possibile solo se si usavano sonde di almeno 600b. Per i geni in singola copia venivano combinati i dati di almeno 30 determinazioni, contando i singoli grani di argento depositati su ogni cromosoma dell'intero genoma, al fine di valutare le zone specifiche di ibridazione ed il background (grani precipitati casualmente su i cromosomi).
Attualmente la tecnica dell'ibridazione *in situ* è migliorata moltissimo con l'uso di sonde legate covalentemente a sostanze fluorescenti. Queste sonde permettono di evidenziare più rapidamente anche geni aventi i loro loci molto vicini (5-700kb). Questa tecnica, indicata con l'acronimo FISH (Fluorescent In Situ Hybridization), non ha sostituito completamente la tecnica AGCSI perché utilizza sonde relativamente grandi, mentre con la tecnica AGCSI che utilizza la PCR analitica si possono mappare rapidamente sequenze di piccole dimensioni come quelle dei marcatori microsatelliti che non è possibile marcare con il FISH.

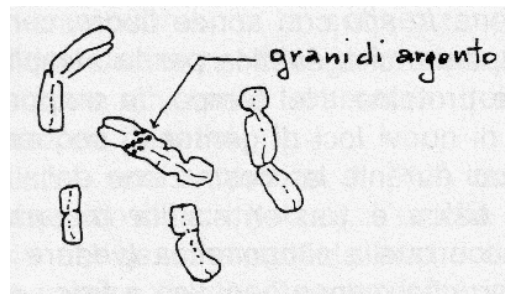



Figura 3-5. Ibridazione *in situ* di un gene in singola copia.

<u>Separazione di cromosomi fluorescenti mediante selezionatore automatico.</u>
Questa macchina detta FACS (Fluorescence-activated cell sorter) è usata per selezionare e quindi concentrare cromosomi omologhi (es. Chr 1 umano) provenienti da cellule di uno stesso tipo (es. cellule coltivate bloccate in metafase con inibitori della mitosi). In questo modo, si può raccogliere un numero sufficientemente grande di copie di uno stesso cromosoma da cui poter estrarre il DNA da analizzare con il metodo Southern o PCR. L'analisi effettuata per ogni cromosoma del genoma umano indica su quale cromosoma sia localizzato il gene od altro tipo di sequenza di interesse.

Costruzione della mappa citogenetica umana

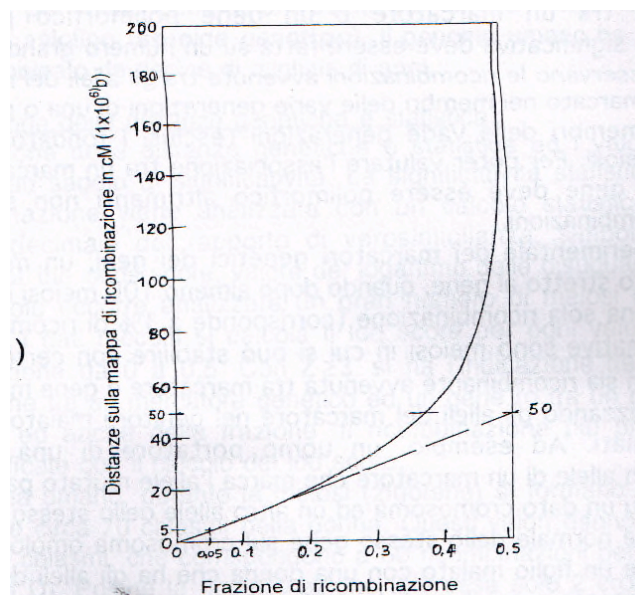
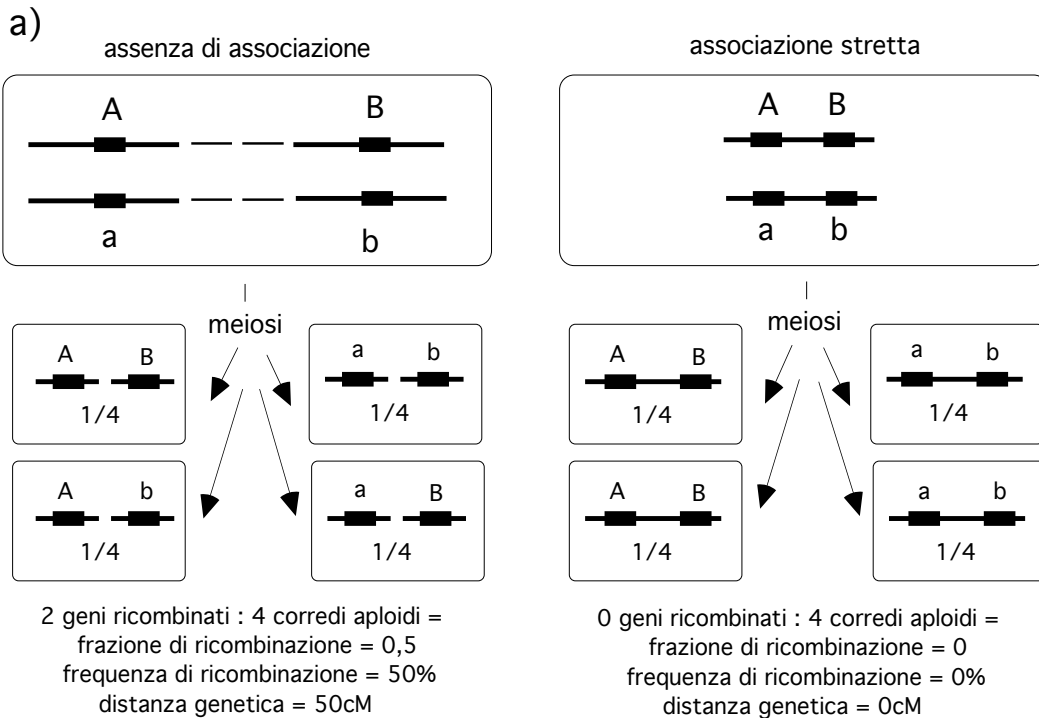
La <u>mappa citogenetica</u> è la descrizione della posizione dei geni o di altre sequenze sulla riproduzione grafica dei cromosomi metafasici colorati.
Il <u>marcatore citogenetico</u> è una sequenza di DNA della quale è nota la posizione su un'unica banda di un dato cromosoma metafasico.
Un gene o qualsiasi altro frammento di DNA può essere mappato citogeneticamente utilizzando la tecnica delle cellule somatiche ibride interspecifiche o l'ibridazione *in situ,* quando sia possibile costruire sonde fluorescenti di almeno 600b. Inoltre geni e frammenti di DNA sono mappati citogeneticamente se si dimostra che la loro sequenza include la sequenza di marcatore citogenetico.
Per la costruzione della mappa citogenetica, i cromosomi metafasici vengono colorati al fine di rivelare bande scure di diversa intensità e bande chiare, rispettivamente regioni più e meno condensate del cromosoma, visibili anche al microscopio ottico (figura 3-14). Queste bande sono numerate dal centromero verso i telomeri con il prefisso p (petit) per il braccio più corto e q per il braccio più lungo. La posizione delle bande serve di riferimento per riconoscere i singoli cromosomi, per confrontare tra loro le mappe citogenetiche di individui della stessa specie, per individuare le aberrazioni cromosomiche, per definire la posizione dei geni ed altre sequenze marcatrici sulla mappa citogenetica. Inizialmente la mappatura dei geni era limitata a quei geni la cui proteina era stata individuata ed essa era facilmente analizzabile utilizzando la tecnica delle cellule somatiche e l'ibridazione *in situ* con sonde radioattive. L'avvento delle

tecnologie del DNA e la scoperta delle molte sequenze non codificanti (esempio: VNTR) disperse nel DNA dei cromosomi, la tecnica delle cellule somatiche ibride e l'ibridazione *in situ* con sonde fluorescenti hanno dato una forte accelerazione alla mappatura citogenetica per la semplicità dell'analisi del DNA rispetto all'analisi delle proteine. Nel tempo, la mappa citogenetica si è progressivamente arricchita di nuovi loci di geni e di sequenze non codificanti ripetute nel genoma individuati durante la costruzione della mappa genetica di associazione e di quella fisica e trasferiti sulla mappa citogenetica per confronto di queste mappe con quella citogenetica (vedere dopo). Quando un nuovo gene viene mappato su una mappa genetica o fisica e su quelle mappe è posto tra due geni già mappati vicini anche sulla mappa citogenetica, confrontando le mappe tra loro si ha l'indicazione del locus del nuovo gene sulla mappa citogenetica (figura 3-14). La mappa citogenetica è classificata come mappa fisica perché le dimensioni delle bande e loro distanze dal centromero sono dimensioni fisiche (es. μm). La dimensione in numero di basi del DNA contenuto nelle bande varia data la diversa condensazione dei cromosomi che comunque è molto alta, pertanto il DNA contenuto nelle bande è approssimativamente dell'ordine di alcuni Mb e questo valore è anche il valore della risoluzione della mappa (tabella 3-3).

Costruzione delle mappe genetiche umane

La prima mappa genetica fu costruita nei primi anni del '900 da Thomas Hunt Morgan ed allievi per il genoma di drosofila utilizzando varianti genetiche visibili morfologicamente. La mappa genetica è chiamata anche mappa genetica di ricombinazione (più raramente di associazione) perché essa è costruita basandosi sulla frequenza di ricombinazione genetica omologa (crossing over) tra sequenze polimorfiche marcatrici del genoma. L'associazione (linkage) tra alleli di geni diversi osservata per più generazioni, cioè per più meiosi, nei membri di una stessa famiglia contrasta con la 2a legge di Mendel che recita: alleli di geni diversi (loci diversi) alla meiosi segregano indipendentemente (appendice D). La legge è vera quando i due geni hanno i loro loci su cromosomi diversi e segregando vanno in gameti diversi (Mendel per caso analizzò solo geni posti su cromosomi diversi e stabilì la legge). Sebbene non detto da Mendel, si potrebbe dedurre che alleli di geni diversi posti sullo stesso cromosoma non segreghino mai. Tuttavia nell'uomo durante la meiosi, avvengono sempre tra i cromosomi omologhi una o più ricombinazioni genetiche omologhe, pertanto geni diversi posti uno vicino ad un telomero e l'altro vicino all'altro telomero di uno stesso cromosoma ricombinano come se fossero su cromosomi diversi (obbedendo così alla 2° legge di Mendel). Tuttavia questo è il caso estremo, perché con il diminuire della distanza fisica (numero di basi) tra i due geni, progressivamente la ricombinazione diviene meno frequente fino ad essere zero per i geni posti fisicamente molto vicini sullo stesso cromosoma. La mappa genetica umana, cioè la mappa di ciascun cromosoma umano, è stata realizzata stabilendo la frequenza di ricombinazione

esistente tra una sequenza polimorfica di riferimento, posta vicino al telomero del braccio p del cromosoma, e le sequenze polimorfiche ed uniche nel genoma, presenti in loci diversi nel DNA dello stesso cromosoma.
Queste sequenze sono chiamate marcatori della mappa genetica o più semplicemente marcatori genetici.
Il marcatore genetico è una sequenza di DNA polimorfica che identifica una unica posizione subcromosomica (locus). L'unicità di locus rende la sequenza del marcatore unica nel genoma aploide anche se il marcatore esiste in più forme diverse in sequenza (alleli). Questa definizione di marcatore genetico è la definizione attuale che include oltre ai geni polimorfici anche le sequenze polimorfiche non codificanti (tabella 3-2) che sono ripetute nel genoma e per questo hanno permesso la costruzione di mappe genetiche di associazione dense di marcatori, cioè costituite da molti marcatori genetici disposti sul DNA di tutti i cromosomi. Un marcatore genetico è marcatore di un gene polimorfico quando un allele del marcatore è associato stretto ad un dato allele del gene, cioè i due alleli, del marcatore e del gene, ricombinano raramente (una volta ogni 100 meiosi) o mai, per cui l'allele del marcatore può essere utilizzato per seguire la trasmissione, da genitori a figli, di un dato gene.
Un marcatore genetico è marcatore di una patologia (anche se il gene responsabile della suscettibilità alla patologia è ancora ignoto) quando un allele del marcatore è associato stretto o presente solo nei pazienti, cioè associato stretto all'allele ignoto responsabile della patologia.
Prima della messa in opera delle tecnologie del DNA, i marcatori genetici erano solamente i geni evidenziabili mediante analisi del loro fenotipo, successivamente con l'avvento delle tecnologie del DNA, la denominazione di marcatore genetico è stata estesa a qualsiasi sequenza polimorfica (anche per una singola base) che identifichi un unico locus. Queste sequenze includono geni e sequenze non codificanti (RFLP, VNTR e SNP).
Il marcatore genetico deve essere polimorfico perché deve marcare con due suoi alleli diversi le molecole di DNA dei due cromosomi omologhi presenti nel genoma nucleare diploide di ogni individuo.
Durante la meiosi, i cromosomi sono duplicati formando due cromatidi fratelli che rimangono uniti per il centromero, quindi i cromosomi omologhi duplicati si appaiano formando una tetrade (complesso di 4 cromatidi omologhi) e due cromatidi non-fratelli (uno di origine paterna ed uno di origine materna) ricombinano, mentre gli altri due cromatidi non ricombinano in quella stessa meiosi. I 4 cromatidi si ripartiranno casualmente in 4 gameti e solo due gameti avranno i cromatidi ricombinati. Questo avviene per ogni tetrade di cromosomi omologhi, così il corredo aploide di ogni gamete includerà cromosomi ricombinati e non ricombinati, in combinazioni che si sono formate casualmente.
Per poter osservare una ricombinazione occorrono due marcatori polimorfici, uno Ab avente il proprio locus a monte del chiasma (es. verso il telomero del braccio piccolo del cromosoma) e l'altro marcatore Bb, a valle del chiasma. Assumiamo che su un cromosoma si trovino gli alleli A-B e sul suo omologo gli

alleli a-b. La ricombinazione è verificata osservando le posizioni relative degli alleli dei due geni. Se assumiamo che gli alleli del marcatore, posto a monte del chiasma, siano rimasti disposti nei due cromosomi omologhi come lo erano prima della meiosi (Aa), gli alleli dell'altro marcatore, posti a valle del chiasma, saranno reciprocamente scambiati in bB. Dopo ricombinazione i cromosomi saranno A-b e a-B ed al termine della meiosi si troveranno in gameti diversi (figura 3-6a). Al termine di ogni meiosi, con una ricombinazione tra due marcatori, 2 gameti avranno un corredo cromosomico con gli alleli dei marcatori ricombinati e gli altri 2 gameti avranno un corredo cromosomico non ricombinato (come prima della meiosi). 2 alleli ricombinati fratto 4 (totale degli alleli formati) è uguale a 0,5 (figura 3-6a). Questo valore è definito <u>frazione di ricombinazione</u> e 0,5 è il suo valore massimo che corrisponde ad una frequenza di ricombinazione del 50%. 50% è la frequenza di ricombinazione di due marcatori posti agli estremi di uno stesso cromosoma perché, durante ogni meiosi, in ogni coppia di cromosomi omologhi, statisticamente avviene almeno una volta una ricombinazione (figura 3-6a).
Lo stesso valore di ricombinazione massima (50%), lo hanno anche i marcatori posti su cromosomi diversi perché durante la meiosi i cromatidi sono distribuiti casualmente nei 4 gameti.
Quando due marcatori posti sullo stesso cromosoma non ricombinano mai, dopo la meiosi si formeranno 4 gameti, ciascuno dei quali avrà 0/4 = 0 ricombinazioni. La frazione di ricombinazione è zero e la frequenza di ricombinazione è uguale a 0% nella regione di DNA che include i loci dei due marcatori. In altre regioni dello stesso cromosoma la ricombinazione può avere valori diversi. Pertanto due marcatori genetici posti sullo stesso cromosoma, in relazione alla distanza fisica tra i loro loci, hanno il valore della loro frazione di ricombinazione compreso tra i valori di 0 e 0,5, rispettivamente corrispondenti a 0 e 50% di frequenza di ricombinazione.
Durante la meiosi, in relazione diretta alla lunghezza dei cromosomi umani, si hanno da 1 a 6 ricombinazioni omologhe per cromosoma, pertanto maggiore è la distanza tra due marcatori genetici posti sullo stesso cromosoma, maggiore è la probabilità che nella stessa meiosi avvengano due ricombinazioni su uno stesso cromosoma. Le due ricombinazioni non avvengono necessariamente nello stesso punto (chiasma) del cromosoma, tuttavia se i due marcatori sono uno a monte e l'altro a valle della regione cromosomica in cui sono avvenuti i due chiasmi, risulta come se non fosse avvenuta nessuna ricombinazione e ciò causa la riduzione della frazione di ricombinazione tra i due marcatori (figura 3-6b).
L'associazione tra un marcatore e un gene polimorfico per essere statisticamente significativa deve essere fatta su un numero grande di meiosi. Per questo si osservano le ricombinazioni avvenute tra gli alleli del marcatore e quelli del gene marcato nei membri delle varie generazioni di una o più famiglie. Il numero dei membri delle varie generazioni (esclusi i fondatori) indica il numero delle meiosi. Per poter valutare l'associazione tra un marcatore ed un

3-6. a) L'associazione tra due geni o marcatori genetici dipende dalla frequenza di ricombinazione che varia da completa indipendenza (frazione di ricombinazione = 0,5, A e B sono su cromosomi diversi o vicini ai due telomeri di uno stesso cromosoma) ad associazione stretta (frazione di ricombinazione = 0, A e B sono fisicamente vicini sullo stesso cromosoma) (da Kaplan J.K. and Delpech (1995) Biologia Molecolare e Medicina, Gnocchi Ed., ridisegnato e modificato). b) Relazione tra frazione di ricombinazione e distanza fisica tra due loci. La relazione lineare è persa per la presenza di più di una ricombinazione per cromosoma che restaura la disposizione originale di tratti dello stesso cromosoma (Connor J.M. and Fergusson-Smith M. A. (1991) Essential Medical Genetics, 3rd ed.).

gene, anche il gene deve essere polimorfico altrimenti non si possono osservare le ricombinazioni.
Nella ricerca sperimentale dei marcatori genetici dei geni, un marcatore è definito associato stretto al gene, quando dopo almeno 100 meiosi informative si è osservata una sola ricombinazione (corrispondente a 1% di ricombinazione). Le meiosi informative sono meiosi in cui si può stabilire con certezza se un gamete sia portatore di una ricombinazione avvenuta tra marcatore e gene marcato. Ciò è verificato analizzando gli alleli del marcatore nel genitore malato e nei suoi discendenti malati. Ad esempio, un uomo portatore di una patologia dominante, ha un allele di un marcatore che marca l'allele mutato patologico di un gene posto su un dato cromosoma ed un altro allele dello stesso marcatore che marca l'allele normale dello stesso gene sul cromosoma omologo. Quest' uomo concepisce un figlio malato con una donna che ha gli alleli dello stesso marcatore dello stesso gene diversi da quelli del futuro padre (ciò permette di distinguere nel figlio i cromosomi del padre da quelli della madre, figura 3-7a). In questa famiglia si può stabilire se la meiosi è informativa perché, se non c'è stata ricombinazione nella meiosi del padre, la posizione degli alleli del marcatore e del gene marcato rimangono come erano, mentre una ricombinazione scambierà gli alleli del marcatore rispetto a quelli del gene di interesse (l'allele che marcava l'allele patologico ora marca quello normale e l'allele che marcava l'allele normale ora marca quello patologico). Assumiamo ora che il padre malato abbia lo stesso allele del marcatore che marca sia l'allele normale che l'allele mutato patologico del gene di interesse. Quando questo padre genererà un figlio malato, non sarà possibile stabilire se l'allele patologico sia stato trasmesso con o senza ricombinazione perché gli alleli del marcatore sono indistinguibili, quindi non informativi (figura 3-7b), pertanto la meiosi è definita non informativa. Egualmente la meiosi non è informativa quando la madre sana abbia ambedue gli alleli del marcatore identici a quello che nel padre marca la patologia.
La marcatura delle patologie genetiche è possibile, anche se il gene mutato che la causa è ignoto, perché di quel gene esistono almeno due alleli: uno normale ed uno mutato responsabile della patologia, individuabili osservando i membri sani e malati di una stessa famiglia. Analizzando l'allele del marcatore genetico associato alla patologia si segue la trasmissione, dai genitori ai figli, dell'allele responsabile della patologia e si può verificare la sua frequenza di ricombinazione osservando la frequenza con la quale l'allele del marcatore è associato ai malati.
Quando gli alleli di due geni, posti sullo stesso cromosoma, non hanno ricombinato per molte generazioni ed in più di 100 meiosi informative, essi formano un aplotipo (**aplo**ide geno**tipo**). Il genoma umano ha aplotipi che non hanno ricombinato da decine di migliaia di anni.

La valutazione delle distanze genetiche è statistica.

La valutazione delle distanze genetiche è statistica ed i valori ottenuti sono sottoposti ad saggio di significatività. La significatività statistica della frazione

di ricombinazione viene analizzata con un calcolo statistico che utilizza il logaritmo decimale del rapporto di verosimiglianza, simbolo Z (lod score, logarithm of the odds score, valore del logaritmo delle disuguaglianze).
Per il calcolo occorre analizzare un gran numero di meiosi (almeno 100) ed utilizzando i dati raccolti si calcola il lod score per ogni valore di frazione di ricombinazione da 0 a 0,5. Con $Z \geq 3$ si ha l'indicazione dell'esistenza della associazione tra un marcatore genetico ed un gene (o tra un marcatore ed una patologia) ed anche della frazione di ricombinazione più probabile tra tutte quelle verificate con il calcolo del lod score.
Nel genoma umano, durante la meiosi (zigotene) si formano circa 55 chiasmi nell'uomo e circa 70 chiasmi nella donna. Questi valori danno un valore medio di circa 60 chiasmi, che corrispondono a 60 ricombinazioni per genoma diploide (appendice D). Poiché la ricombinazione interessa solo 2 cromatidi non-fratelli su 4 cromatidi omologhi, si ottiene una media di 30 chiasmi per coppia di cromatidi, quindi 30 ricombinazioni per genoma (2N). Poiché le ricombinazioni avvengono tra cromosomi omologhi, anche i cromosomi del genoma aploide avranno subito mediamente 30 ricombinazioni. Pertanto il numero delle ricombinazioni del corredo N è uguale a quello 2N.
Ad un chiasma, che è l'immagine della ricombinazione, è stato dato il valore di 1 Morgan, cioè 100% di ricombinazione perché sono considerati solo gli alleli sui quali sono avvenuti i 60 chiasmi ed essendo andati tutti incontro a ricombinazioni, la frequenza di ricombinazione totale è 100%.
Il **Morgan (M)** è l'unità di misura della distanza genetica (in onore del genetista T.H. Morgan).
La distanza genetica massima e quella minima tra un marcatore ed un gene e tra qualsiasi altro tipo di sequenza sono rispettivamente 50M e 0M (figura 3-6b e Tabella 3-1).
Abbiamo visto che, ad ogni meiosi, il genoma aploide partecipa a 30 chiasmi a cui corrispondono 30M = 3.000cM.
La distanza genetica massima e quella minima tra un marcatore ed un gene e tra qualsiasi altro tipo di sequenza sono rispettivamente 50M e 0M (figura 3-6b e Tabella 3-1).

Tabella 3-1. X = non misurabile.

Distanza fisica b	Distanza genetica cM	Frequenza di ricombinazione %	Frazione di ricombinazione
10M	10	10	0,10
5M	5	5	0,05
1M	1	1	0,01
100k	0,1	0,1	0,001
10k	x	x	x
1k	x	x	x

Dati da Kaplan J.K. and Delpech M.(1995) Biologia Molecolare e Medicina, Gnocchi Ed., Napoli, in parte modificati ed altri aggiunti.

Abbiamo visto che, ad ogni meiosi, il genoma aploide partecipa a 30 chiasmi a cui corrispondono 30M = 3.000cM.
Dividendo $3x10^9$b (numero di basi del genoma aploide umano) per 3.000cM si ottiene un milione di basi per ogni cM e cioè **1cM corrisponde a 1Mb.**
La distanza genetica non è una lunghezza fisica, cioè non corrisponde ad un definito numero di coppie di basi, tuttavia le distanze genetiche tra marcatori sono in relazione diretta con la distanze fisiche (numero di basi) tra i marcatori, nel senso che maggiore è il numero dei cM esistenti tra due geni e maggiore è il numero di basi che separa i due geni. Ciò è dimostrato dalla esatta corrispondenza dell'ordine dei loci nelle mappe genetiche con quello dei loci delle mappe fisiche (figura 3-14). Tuttavia la corrispondenza riguarda l'ordine della sequenza dei loci sul cromosoma e non le distanze genetiche con quelle fisiche tra i marcatori. Infatti, la corrispondenza tra le unità di misura cM e Mb è molto approssimativa, per il fatto che in uno stesso cromosoma esistono regioni di DNA che ricombinano con frequenze molto diverse, e anche il locus di ricombinazione non è casuale.
I cromosomi umani sono costituiti da blocchi di circa 20-50kb geneticamente conservati separati da più piccoli blocchi di 1-2kb, detti punti caldi (hot spot), ed il 95% di tutte le ricombinazioni avviene in questi punti caldi. Esistono regioni cromosomiche lunghe fino a 5Mb, chiamate deserti di ricombinazione, dove due geni possono essere distanti fisicamente 1Mb (che mediamente corrisponde a 1M) mentre la distanza genetica è meno di 0,3cM (3 ricombinazioni ogni 1000 meiosi). Mentre nelle regioni cromosomiche, dette giungle di ricombinazione, due geni distanti fisicamente 1Mb distano geneticamente più di 3cM (3 ricombinazioni ogni 100 meiosi).
La ricombinazione inizia sempre su una sequenza di circa 10bp, tuttavia inspiegabilmente questa sequenza non ha caratteristiche tali da spiegare la specificità del locus di ricombinazione. La frequenza di ricombinazione genica varia anche in relazione al sesso, essa è più alta nelle donne che nell'uomo (anche geneticamente "la donna è mobile" Rigoletto, G. Verdi, libretto F.M. Piave). Nella donna il cromosoma 1 ha una lunghezza genetica di 500cM e nell'uomo di 305cM, mentre la lunghezza fisica (numero di basi) del cromosoma 1 è la stessa nella donna e nell'uomo. La lunghezza genetica del genoma femminile è 4300cM, circa il 51% più lunga di quella del genoma maschile che è 2850cM. Questi sono valori medi perché nelle regioni subtelomeriche ed in quelle con impronta parentale (appendice D) si ha una maggiore ricombinazione nei cromosomi maschili rispetto a quelli femminili, mentre nelle regioni centromeriche si hanno ricombinazioni nelle femmine e non nei maschi. Si è osservato che il genitore che ha il sesso eterogametico (cromosomi diversi), nella specie umana è l'uomo con XY, ricombina meno frequentemente di quello omogametico (femmina: XX) e ciò è stato messo in relazione alla durata maggiore della metafase 1 (dove avviene la ricombinazione) della meiosi femminile. Nonostante queste differenze è accettato il compromesso di 1cM = 1Mb mettendo così d'accordo maschi, femmine, giungle e deserti di ricombinazione.

Quando un marcatore è distante da un gene 1cM (1 ricombinazione su 100 pazienti) è detto associato stretto al gene e ciò garantisce che il marcatore possa essere utilizzato per analisi genetiche del gene marcato. Inoltre 1cM corrisponde ad una distanza fisica di 1Mb, che sebbene sia approssimativa, è considerata sufficientemente piccola per intraprendere la ricerca del gene, determinando la sequenza del DNA che nello stesso cromosoma separa il locus del marcatore da quello del gene (vedere capitolo 4).

Ricerca dei marcatori genetici

Geni marcatori genetici.
Prima dell'avvento delle tecnologie del DNA, l'unica forma possibile di marcatore genetico erano i geni polimorfici che codificavano una proteina di facile dosaggio ed essi erano utilizzati principalmente per marcare malattie genetiche (esempio, tipi dei globuli rossi, figura 3-7).
Un gene può essere usato come marcatore genetico di una malattia genetica anche quando il gene responsabile della malattia può essere ignoto ed ignota la proteina da esso codificata.
Il gene marcatore di un gene deve essere polimorfico, esistere in almeno due forme alleliche ed essere strettamente associato al gene mutato responsabile della malattia. Un dato allele del gene marcatore deve essere presente nel DNA dello stesso cromosoma dove si trova l'allele mutato patologico del gene della suscettibilità alla patologia e quindi presente solo nei portatori della patologia e mai in individui sani della stessa famiglia. L'altro allele del gene marcatore deve esse sempre presente sul cromosoma, omologo di quello patologico, dove si trova l'allele sano del gene della suscettibilità alla patologia e quindi presente solo in individui sani e mai su quelli portatori della malattia. Se si realizzano queste condizioni, è sufficiente dimostrare la presenza nel genoma dell'allele del gene marcatore per stabilire che un individuo è portatore dell'allele responsabile della patologia e si dice che quell'allele di quel gene marcatore è informativo (o diagnostico) della patologia nella famiglia analizzata.
La ricerca di un marcatore genetico di una patologia può essere fatta senza necessariamente conoscere la posizione cromosomica e subcromosomica del marcatore, è sufficiente aver dimostrato che essi sono geneticamente associati. E' importante che il prodotto degli alleli del gene marcatore della malattia sia analizzabile mediante una metodologia con alto grado di affidabilità, cioè che sia riproducibile nel 100% dei casi.
Individuato il marcatore di un gene patologico, esso è poi utilizzato per scopi diagnostici, per individuare i portatori dell'allele patologico nei membri della stessa famiglia ed in quelli di altre famiglie.
La capacità diagnostica di un marcatore genetico di una data patologia ha l'importante funzione di individuare i portatori del cromosoma patologico prima che la malattia genetica si manifesti alterando l'organismo (sintomi clinici). Ciò risulta di fondamentale utilità per le analisi prenatali di malattie monogeniche dominanti e recessive e per le analisi post-natali delle malattie dominanti che,

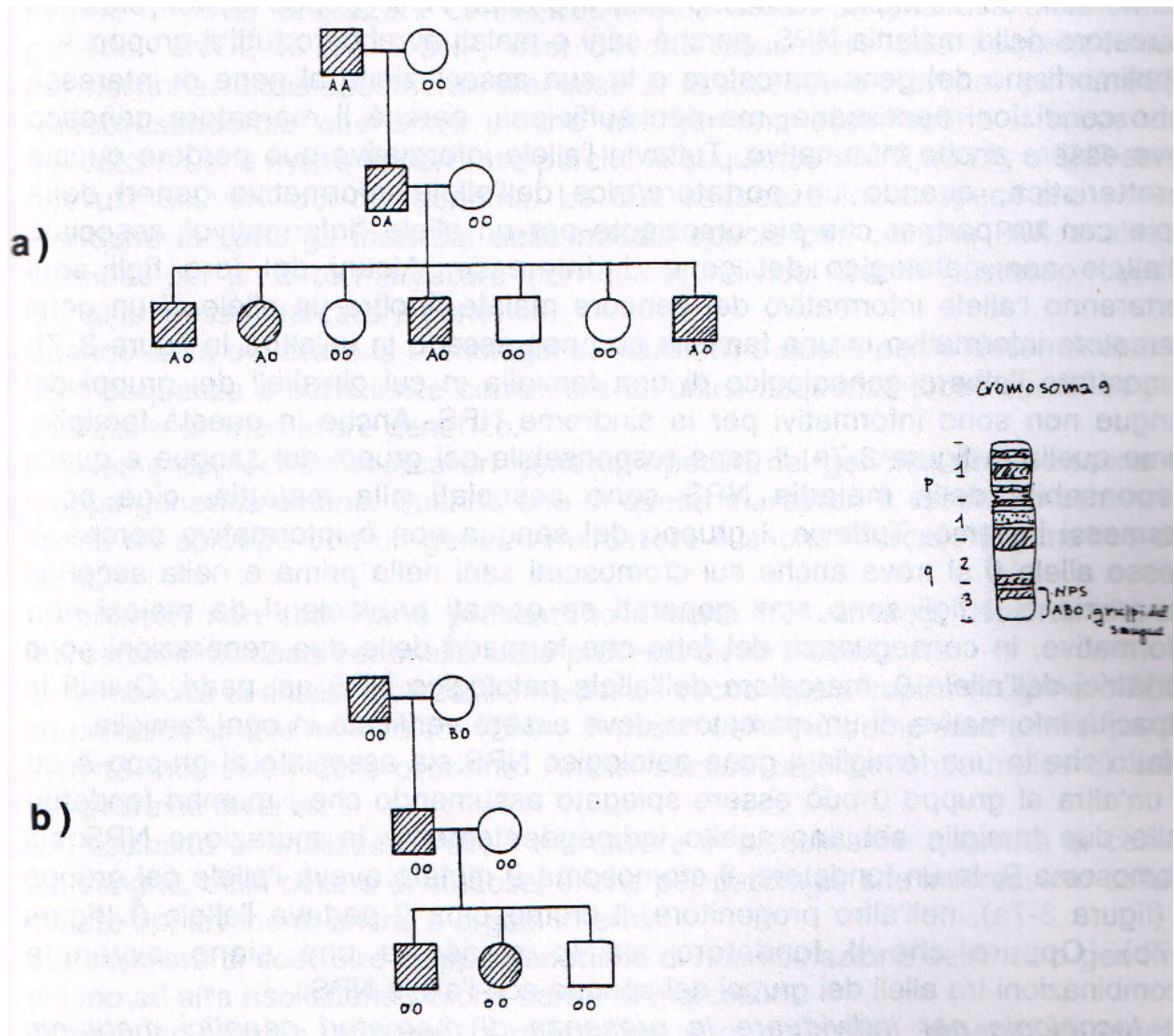



Figura 3-7. Alberi genealogici di due famiglie con membri portatori della sindrome "Unghie-rotula" (NPS) autosomica dominante il cui gene sul cromosoma 9 è associato a quello polimorfico responsabile dei gruppi del sangue. a) il gruppo del sangue (allele A) è informativo perché in questa famiglia chi ha il gruppo del sangue A è anche portatore della sindrome NPS, l'individuo maschio della prima generazione è omozigote per il gene del gruppo del sangue A ed eterozigote per l'allele mutato del gene della suscettibilità alla NPS. Il cromosoma 9 con l'allele mutato è stato trasmesso, in seconda generazione, all'unico figlio maschio, il quale lo ha trasmesso a 4 dei suoi 7 figli; b) in questa famiglia il gruppo del sangue (allele 0) non è informativo perché lo hanno sia i membri sani che quelli portatori della sindrome NPS. Per i simboli ed altri dati vedere figura 3-1 e testo. (ridisegnato e modificato da Connor J.M. and Fergusson-Smith M. A. (1991) Essential Medical Genetics, 3rd ed., Blackwell, London).

per scarsa penetranza, si manifestano in tarda età (appendice E). In quest'ultimo caso, l'assenza di manifestazioni patologiche non condiziona un portatore adulto, che essendo inconsapevole, può decidere di procreare. L'analisi del marcatore genetico è utile anche per individuare adulti portatori sani (eterozigoti) di malattie recessive al momento che si scopre che altri membri della stessa famiglia sono portatori della patologia.
Un esempio di gene marcatore è il gene polimorfico responsabile dei tre principali gruppi del sangue (alleli A, B, O), il gene è localizzato sul braccio lungo del cromosoma 9 (Chr.9q) ed è il marcatore genetico della sindrome "Unghie-rotula" (NPS), patologia dominante che si manifesta con l'assenza della rotula, unghie distrofiche e disturbi renali (figura 3-7).
Quando tra i membri di una stessa famiglia i portatori della patologia NPS hanno, ad esempio, tutti il gruppo del sangue A (figura 3-7a) e quelli sani il gruppo del sangue 0, si deduce che i due geni (il gene della suscettibilità alla NPS ed il gene responsabile dei gruppi del sangue) siano associati, cioè nelle generazioni considerate non ci sono stati eventi di ricombinazione omologa.
Quindi il gene responsabile dei gruppi del sangue può essere definito marcatore del gene responsabile della sindrome NPS.
Il polimorfismo del gene responsabile dei gruppi del sangue è fondamentale per seguire la malattia NPS tra i membri della famiglia perché, senza conoscere il gene responsabile della sindrome NPS, si può stabilire che in quella famiglia i nascituri con gruppo del sangue A saranno malati NPS e quelli con gruppo 0 sani. Se il gene responsabile dei gruppi del sangue nell'uomo non fosse polimorfico, ad esempio se esistesse solo il gruppo A, il gruppo A non sarebbe marcatore della malattia NPS, perché sani e malati avrebbero tutti il gruppo A.
Il polimorfismo del gene marcatore e la sua associazione al gene di interesse sono condizioni necessarie, ma non sufficienti, perché il marcatore genetico deve essere anche informativo. Tuttavia l'allele informativo può perdere questa caratteristica, quando un portatore/trice dell'allele informativo generi della prole con un partner che sia omozigote per un allele "informativo" associato all'allele non patologico del gene di interesse. Alcuni dei loro figli sani porteranno l'allele informativo del genitore malato. Inoltre un allele di un gene marcatore informativo in una famiglia può non esserlo in un'altra. In figura 3-7b è mostrato l'albero genealogico di una famiglia in cui gli alleli dei gruppi del sangue non sono informativi per la sindrome NPS. Anche in questa famiglia, come quella in figura 3-7a, il gene responsabile dei gruppi del sangue e quello responsabile della malattia NPS sono associati alla malattia, cioè sono trasmessi insieme. Tuttavia il gruppo del sangue non è informativo perché lo stesso allele 0 si trova anche sui cromosomi sani nella prima e nella seconda generazione. I figli sono stati generati da gameti provenienti da meiosi non informative, in conseguenza del fatto che le madri delle due generazioni sono portatrici dell'allele 0, marcatore dell'allele patologico NPS nei padri. Quindi la capacità informativa di un marcatore deve essere verificata in ogni famiglia.
Il fatto che in una famiglia il gene patologico NPS sia associato al gruppo A ed in un'altra al gruppo 0 può essere spiegato assumendo che i membri fondatori

delle due famiglie abbiano subito indipendentemente la mutazione NPS sul cromosoma 9. In un fondatore, il cromosoma 9 mutato aveva l'allele del gruppo A (figura 3-7a), nell'altro progenitore, il cromosoma 9 portava l'allele 0 (figura 3-7b). Oppure che il fondatore sia lo stesso e che siano avvenute ricombinazioni tra alleli dei gruppi del sangue con l'allele NPS.

Le tecnologie per individuare la presenza di marcatori genetici mediante l'analisi dell'attività o delle caratteristiche molecolari delle proteine sono state progressivamente abbandonate con l'avvento delle moderne tecnologie per l'analisi diretta del DNA (tecnologia di Southern, della PCR e dell'analisi della sequenza). Queste tecnologie hanno permesso di individuare e clonare altri tipi di marcatore genetico polimorfico costituito da sequenze non codificanti ripetute e disperse nel genoma e dati i numerosi loci di questi marcatori, è stato possibile costruire dettagliate mappe genetiche del genoma umano includenti decine di migliaia di marcatori.

I <u>nuovi marcatori genetici</u> sono le sequenze non codificanti polimorfiche RFLP, VNTR e SNP, che sono ripetute nel genoma e le varie copie delle sequenze sono poste su loci diversi (tabella 3-2). Queste sequenze possono avere il loro locus all'interno della sequenza dei geni. Le sequenze RFLP e SNP possono essere all'interno di introni ed esoni, mentre le sequenze VNTR sono quasi esclusivamente all'interno di introni. Le sequenze RFLP, VNTR e SNP contribuiscono al polimorfismo dei geni e possono essere responsabili di proprietà minori, subdole o patologiche (appendici D ed E).
Al fine di poter analizzare con la tecnologia della PCR analitica i marcatori genetici VNTR, RFLP, o SNP, essi devono possedere due sequenze non polimorfiche, locus-specifiche alle quali si assoceranno i primer per la PCR. Queste sequenze che sono uniche nel genoma stabiliscono il locus dei marcatori. Ciò è molto importante perché le sequenze VNTR, RFLP, o SNP sono ripetute su vari loci del genoma. Le due sequenze locus specifiche sono identiche in tutti gli individui della nostra specie per cui una PCR analitica specifica per un dato marcatore individua in individui diversi lo stesso locus e quindi lo stesso marcatore genetico.
Quando sono utilizzate la tecnologia di Southern o quella per la determinazione della sequenza è sufficiente conoscere un'unica sequenza locus specifica per analizzare un marcatore genetico.
L'insieme dei loci dei marcatori genetici ripetuti nel genoma costituiscono la mappa genetica umana. Quando uno di questi marcatori è associato stretto o forma un aplotipo con un gene, il marcatore è anche marcatore genetico del gene.
I marcatori non codificanti presentano almeno tre vantaggi rispetto ai geni marcatori individuati per analisi della proteina da loro codificata:
1. Semplicità di analisi molecolare mediante tecnologie semplici (in genere PCR) su un'unica specie molecolare (DNA) al posto delle più complesse analisi per le attività molecolari delle proteine. Analisi spesso tecnicamente molto diverse per proteine diverse.

2. Possibilità di analizzare l'allele marcatore in piccolissime quantità di cellule del sangue, della pelle o di mucose, anche per patologie che interessano cellule malate appartenenti anche a organi interni.
3. Possibilità di costruire mappe genetiche di ricombinazione dell'intero genoma umano ad alta risoluzione perché dense di marcatori.
Dopo aver clonato un gene, la ricerca dei suoi marcatori genetici in genere viene fatta verificando la presenza di sequenze di microsatelliti all'interno di introni ed estendendo l'analisi della sequenza alle regioni al 5' ed al 3' del gene per circa 1-2Mb. Viene verificato il polimorfismo della sequenza VNTR e ricercate le sequenze locus specifiche.
Data la ripetitività della loro sequenza, individuare le sequenze VNTR (mini e microsatelliti) in una sequenza di DNA è più semplice che individuare le sequenze RFLP e SNP. Per stabilire che siano sequenze marcatrici genetiche occorre verificare il loro polimorfismo, in genere le sequenze VNTR hanno un grado di polimorfismo superiore a quello degli altri due tipi di marcatore che può essere solo di 2 e 4 alleli rispettivamente per le sequenze SNP e RFLP.
Per rendere possibile e specifica l'analisi del DNA di un marcatore genetico polimorfico non codificante proteine occorre:

1. Clonare e determinare la sequenza della regione del DNA genomico che include la sequenza nucleotidica polimorfica del marcatore.
2. Individuare, al 5' ed al 3' della sequenza polimorfica, due sequenze locus-specifiche diverse tra loro e capaci di associare specificamente alla stessa Tm i primer per eseguire la PCR analitica.
3. Analizzare con la tecnica PCR il DNA del marcatore nel genoma di vari individui al fine di stabilire la sua locus specificità ed il suo polimorfismo all'interno di una famiglia o di una popolazione.

La locus specificità di un marcatore è verificata analizzando l'associazione genetica del marcatore con un altro marcatore genetico o gene di noto locus.
L'evoluzione delle tecniche per ricerca dei marcatori è stata:
analisi molecolare di proteine codificate da geni marcatori → analisi Southern sul DNA di marcatori non codificanti proteine → analisi della PCR sul DNA marcatori non codificanti proteine.

Marcatori genetici costituiti da DNA non codificante e ripetuto nel genoma

<u>Marcatori genetici del polimorfismo della lunghezza dei frammenti di restrizione.</u>
Sono chiamati marcatori genetici RFLP (RFLP = Restriction Fragment Length Polymorphism), oppure marcatori genetici RSP (Restriction Site Polymorphism).
I marcatori genetici RFLP sono i siti di riconoscimento e di taglio degli enzimi di restrizione ed per ogni diverso tipo di enzima hanno molti siti nel DNA genomico nucleare umano. Il polimorfismo di questi siti è di solo 2 alleli e risulta da mutazioni puntiformi della sequenza palindromica di restrizione che mutata non è più riconosciuta dal rispettivo enzima. In quel punto il DNA non sarà più tagliato da quel dato enzima di restrizione, così dopo digestione del DNA,

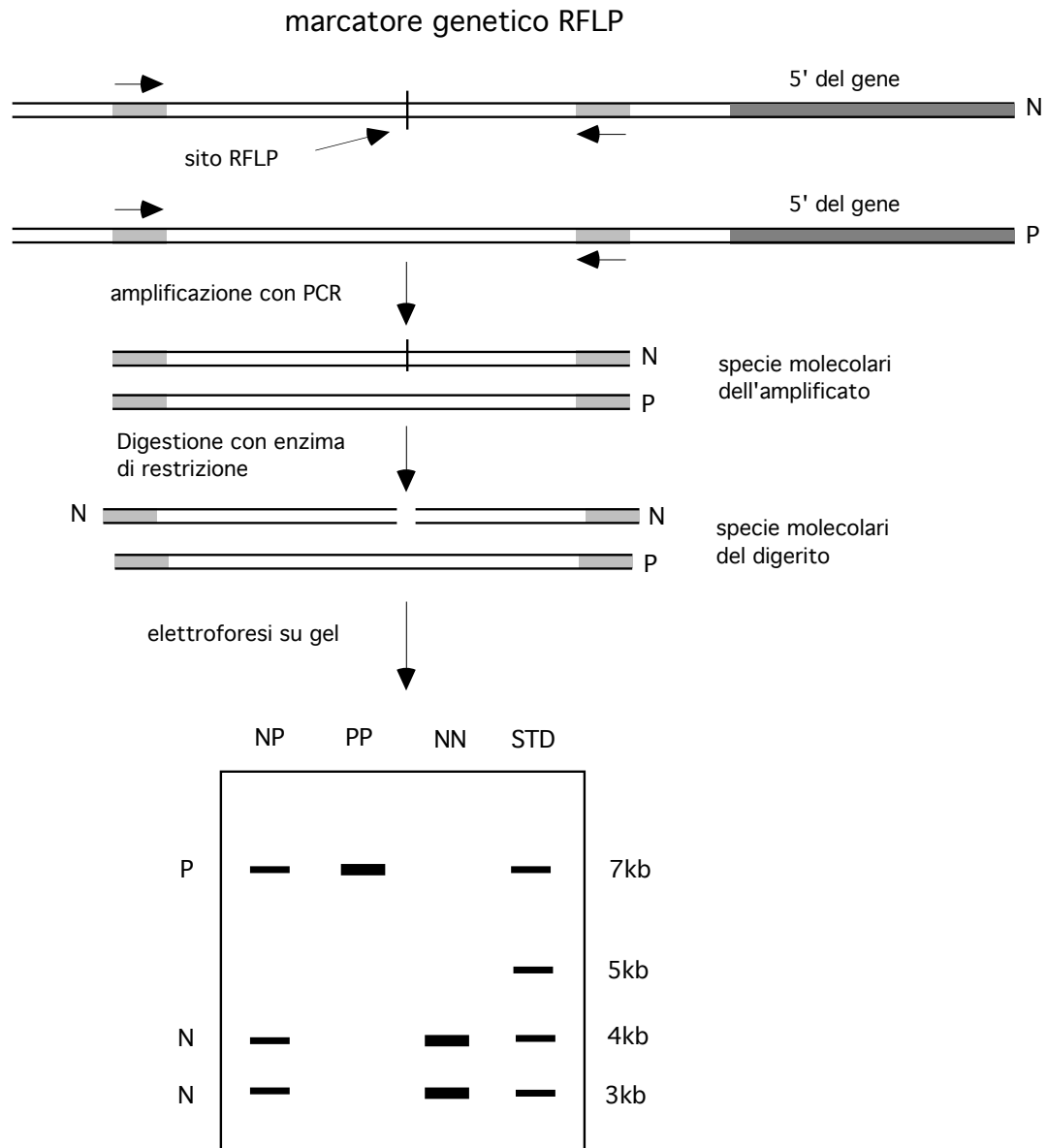



Figura 3-8. Identificazione del marcatore genetico RFLP mediante la tecnica della PCR in condizioni di omozigosi e di eterozigosi.
Il marcatore si trova al 5' del promotore del gene marcato. Le frecce orizzontali rappresentano i primer per la PCR che sono complementari a regioni che identificano il locus del marcatore. N = allele normale, P = allele patologico, STD = marcatori di dimensione (frammenti di DNA di noto numero di basi).

La presenza e l'assenza del sito di restrizione è analizzata amplificando specificamente con la tecnica della PCR la regione includente il sito di restrizione polimorfico, l'amplificato è digerito con l'enzima di restrizione di interesse e poi analizzato mediante elettroforesi. Dal numero dei frammenti (bande elettroforetiche) e dalla lunghezza dei frammenti (numero di basi) si ha l'indicazione della presenza o meno del sito di restrizione (figura 3-8). In condizioni di eterozigosi si avranno 3 bande: una banda contenente il frammento di dimensioni più grandi (in figura, allele P) ed altre due bande includenti frammenti di lunghezza diversa (allele N), la somma delle basi dei quali è uguale al numero di basi del frammento più lungo.
Per poter eseguire la PCR occorre aver clonato in precedenza la regione contenente il sito di restrizione, averne determinato la sequenza ed aver individuato due sequenze specifiche per il locus di quel sito VNTR.
A queste sequenze si assoceranno i primer della PCR sintetizzati complementari a dette regioni.
Un procedimento usato per ricercare un marcatore RFLP che marchi una patologia consiste nell'effettuare, separatamente con enzimi di restrizione diversi, più digestioni del DNA dei membri sani e di quelli malati della famiglia. Poi viene analizzato, mediante Southern, il DNA delle varie digestioni, fino a quando non si trova che un enzima di restrizione, che pur producendo tanti frammenti uguali nei membri sani e nei membri malati, nei malati ne produce uno più lungo che non è mai presente nei sani (oppure il contrario, più lungo nei sani ed assente nei malati). Il frammento di DNA più lungo viene clonato, la sua sequenza determinata e ricercate due sequenze locus specifiche per poterlo poi analizzare con la PCR analitica. Il marcatore RFLP della patologia è individuato quando l'associazione tra l'assenza del sito e la patologia è confermata in almeno 100 meiosi informative.
Conoscere il cromosoma dove si trova il marcatore non è necessario per individuare i portatori della patologia anche in altre famiglie e per analisi prenatali, tuttavia il locus citogenetico del sito RFLP di interesse può essere individuato con PCR analitica facendo un'analisi delle cellule somatiche ibride interspecifiche (figura 3-4). Per i marcatori genetici non codificanti, data la loro piccola dimensione, non può essere utilizzata la tecnica dell'ibridazione *in situ.*
Può accadere che in una famiglia la malattia sia associata al sito di restrizione e la sanità all'assenza del sito, perché non è l'assenza o la presenza del RFLP che è responsabile della patologia. I marcatori di patologie sono come gli ambasciatori: non portano pena, la comunicano solamente.
In una famiglia in cui alcuni membri maschi erano affetti da distrofia muscolare di Ducenne (DMD) si osservò che il gene patologico, che è posto sul cromosoma X, era associato alla presenza di un sito dell'enzima di restrizione BglI (denunciata dalla presenza di due frammenti di restrizione), mentre i maschi sani non avevano tale sito. Con l'analisi Southern fu possibile individuare nella stessa famiglia le femmine portatrici della DMD. Inoltre fu osservato che in alcuni membri della stessa famiglia era presente un cromosoma X sano che aveva il sito BglI marcatore della DMD (in questo caso il marcatore non era

informativo). Questo dato (sito BglI presente nei sani e nei malati) servì a dimostrare che la mutazione responsabile del polimorfismo RFLP associato al gene DMD è solo un marcatore della DMD e non è coinvolta nella patologia perché presente anche in individui sani. Tuttavia si può ipotizzare che un sito RFLP, usato come marcatore, sia all'interno di un esone di un gene e che una sua mutazione causi una patologia, in questo caso il marcatore è anche responsabile di una patologia.
Le sequenze RFLP polimorfiche, essendo distribuite su tutto il genoma, sono state usate per costruire la prima mappa genetica del genoma umano.

Prima dell'avvento della tecnica della PCR, il polimorfismo RFLP era analizzato mediante un'analisi Southern del DNA genomico, digerito con un enzima di restrizione che tagliava il DNA genomico in siti al 5' ed al 3' del sito RFLP polimorfico. Lo stesso enzima tagliava o non tagliava il DNA nel sito polimorfico RFLP. I frammenti di DNA genomico, analizzati con la tecnologia Southern, erano identificati con una sonda specifica per la regione contenente il sito RFLP polimorfico, capace di associarsi sia al frammento intero di DNA (non digerito), che ad ambedue i frammenti di DNA digerito.

Marcatori genetici del Numero Variabile di Ripetizioni in Tandem
I marcatori genetici VNTR (Variable Number of Tandem Repeat) sono costituiti da sequenze VNTR (capitolo 1 e Tabelle 1-2 e 3-2) ed in particolare da minisatelliti e microsatelliti (esclusi i minisatelliti costituiti da ripetizioni di singole basi). Minisatelliti e microsatelliti sono preferiti per il loro alto livello di polimorfismo, per la loro presenza in più loci del genoma umano e perché analizzabili con la tecnica della PCR analitica. Per poterli analizzare con la tecnica PCR la regione del loro locus è stata clonata, sequenziata e sono state identificate le sequenze locus-specifiche poste al 5' ed al 3' della sequenza VNTR. In un dato locus, dato il polimorfismo, le ripetizioni in tandem di un marcatore VNTR possono variare da individuo ad individuo, ma le due sequenze uniche locus-specifiche sono uguali in tutti gli individui umani.
La tecnica della PCR permette di amplificare specificamente la regioni di DNA genomico, locus del marcatore VNTR, e di analizzare le diverse dimensioni degli amplificati mediante elettroforesi (figura 3-9).

Prima dell'avvento della tecnica della PCR, i marcatori VNTR erano analizzati con la tecnologia Southern del DNA genomico, digerito con un enzima di restrizione che tagliava al 5' ed al 3' il DNA includente la sequenza VNTR. La sonda utilizzata ibridava con il DNA di una regione locus-specifica contigua alla sequenza VNTR (figura 3-15).

Nella famiglia di figura 3-9, i membri sani hanno l'allele della suscettibilità alla patologia (sano) associato all'allele marcatore VNTR con 6 ripetizioni in tandem $(AC/GT)_6$, mentre i membri malati hanno l'allele mutato patologico associato all'allele marcatore con 4 ripetizioni in tandem $(AC/GT)_4$. In questa famiglia è possibile seguire la trasmissione dell'allele patologico nelle varie generazioni ed individuare i portatori della patologia analizzando l'allele con 4 ripetizioni in tandem.

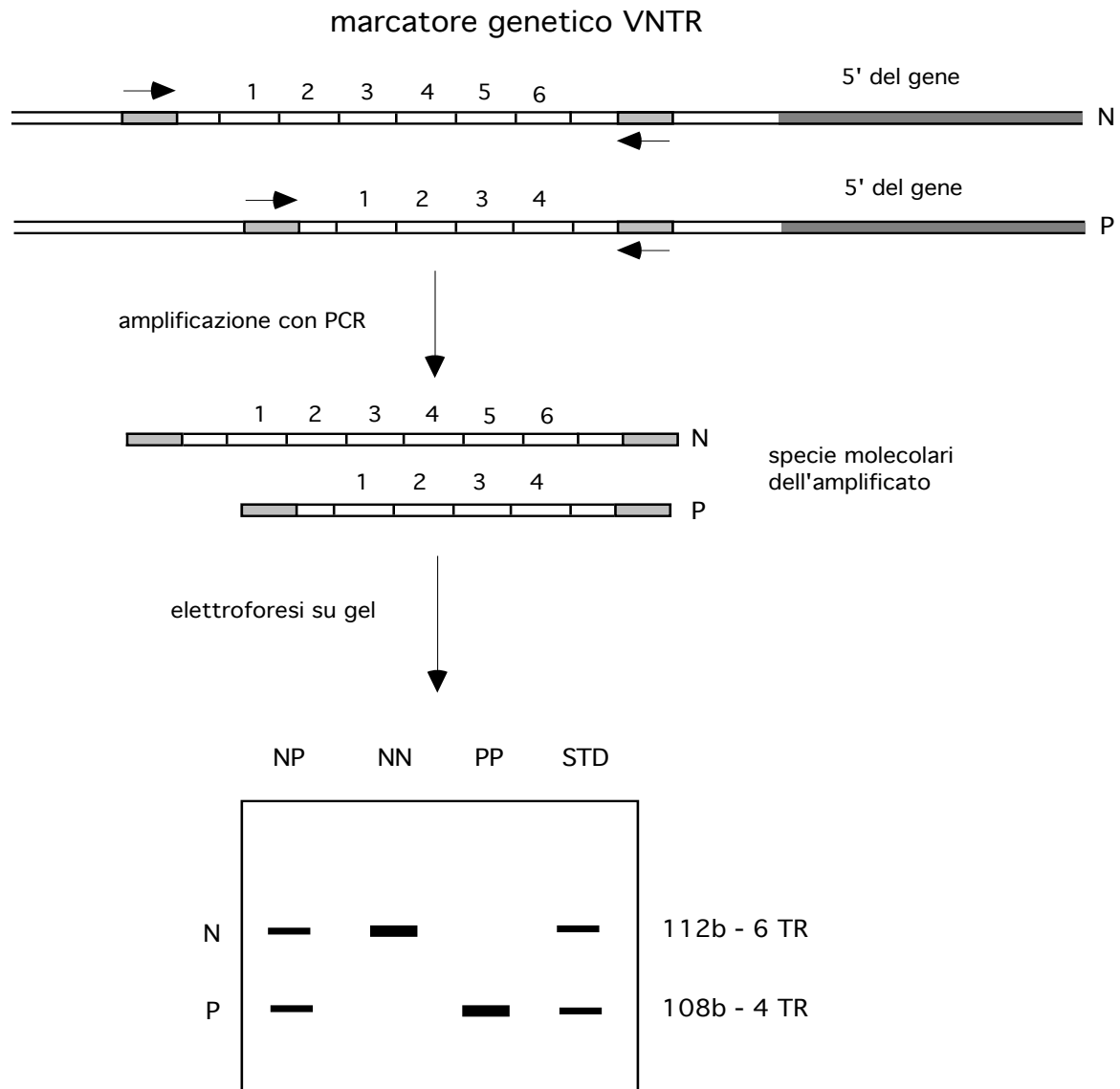



Figura 3-9. Identificazione del marcatore genetico VNTR mediante la tecnica della PCR in condizioni di omozigosi ed eterozigosi. Le frecce orizzontali rappresentano i primer per la PCR che sono complementari a regioni di DNA (rettangoli punteggiati). Queste regioni sono uniche nel genoma ed identificano il locus del marcatore. N = allele normale, P = allele portatore patologico, STD = marcatori di dimensione (frammenti di DNA di noto numero di basi). Le dimensioni dei frammenti di DNA amplificato includono le ripetizioni in tandem (TR) e le regioni ai loro 5’ e 3’ che, in uno stesso locus, hanno lo stesso numero di basi, pertanto si possono valutare le variazioni dei TR mediante una elettroforesi su gel.

Il procedimento, usato per ricercare un marcatore VNTR di una patologia genetica, è simile a quello descritto sopra per i marcatori RFLP. Si operano separatamente varie digestioni con enzimi di restrizione diversi. Poi , mediante Southern, si analizza il DNA delle varie digestioni utilizzando come sonda il DNA della sequenza unitaria di una VNTR e verificando la presenza di frammenti presenti e assenti nel DNA sia dei membri malati che di quelli sani della stessa famiglia (l'autografia ha un immagine simile a quella delle corsie elettroforetiche M ed F della figura 3-15). In caso che non si trovino differenze tra le bande si usa coma sonda il DNA della sequenza unitaria di altre sequenze VNTR, fino a trovare delle differenze sopra indicate.
Ad esempio, individuata una banda presente solo nei membri malati, si clona il suo DNA, si individuano due sequenze uniche nel genoma al 5' e 3' della sequenza VNTR e mediante PCR analitica si verifica che quella sequenza VNTR amplificata sia la stessa (stessa dimensione) in tutti i membri malati e che nei membri sani sia presente la stessa sequenza VNTR con un numero diverso di ripetizioni. Gli stessi primer garantiscono di operare sullo stesso locus, la diversità di dimensione delle bande mostra che il marcatore VNTR della patologia è polimorfico e che un suo allele marca l'allele patologico e l'altro l'allele sano. Per confermare il risultato si analizza il DNA di almeno 100 meiosi informative. Questa stessa analisi permette di stabilire anche il grado di polimorfismo del marcatore e se il marcatore è informativo in tutte le famiglie analizzate. Le sequenze marcatrici minisatellite e microsatellite sono state le più utilizzate per marcare le patologie (vedere capitolo 4) e sono utilizzate in medicina legale per identificare individui. (figura 3-16).

Marcatori del polimorfismo di un singolo nucleotide

Nel genoma umano il polimorfismo di un singolo nucleotide si verifica 1 volta ogni circa 1000 basi (circa 3 milioni di loci nel genoma umano, Tabella 3-2) e crea anche il polimorfismo RFLP. Teoricamente la mutazione puntiforme in una data posizione di una sequenza può generare 3 tipi di varianti SNP della base normale. La sostituzione della base può avvenire per transizione: purina-->purina (es. A-->G) o pirimidina-->pirimidina (es. C-->T) o per transversione: purina-->pirimidina (es. A-->T) o pirimidina-->purina (es. T-->A). In natura per alcuni loci si sono osservate le 4 possibili varianti SNP, tuttavia la condizione più frequente è che le varianti siano solo 2 provocate per lo più per transizione. La spiegazione di questa conservazione è attribuita all’osservazione che, in natura, la mutazione per transizione è più frequente di quella per transversione e che una volta avvenuta una mutazione è improbabile, data la casualità dell’evento, che sulla stessa posizione (locus) avvenga un'altra mutazione nei discendenti del progenitore che ha subito la prima mutazione.
I marcatori SNP (SNP = Single Nucleotide Polymorphism) sono analizzati facendo la sequenza della regione di DNA genomico che include il locus SNP utilizzando un primer complementare ad una sequenza unica nel genoma marcatrice del locus

marcatore genetico SNP

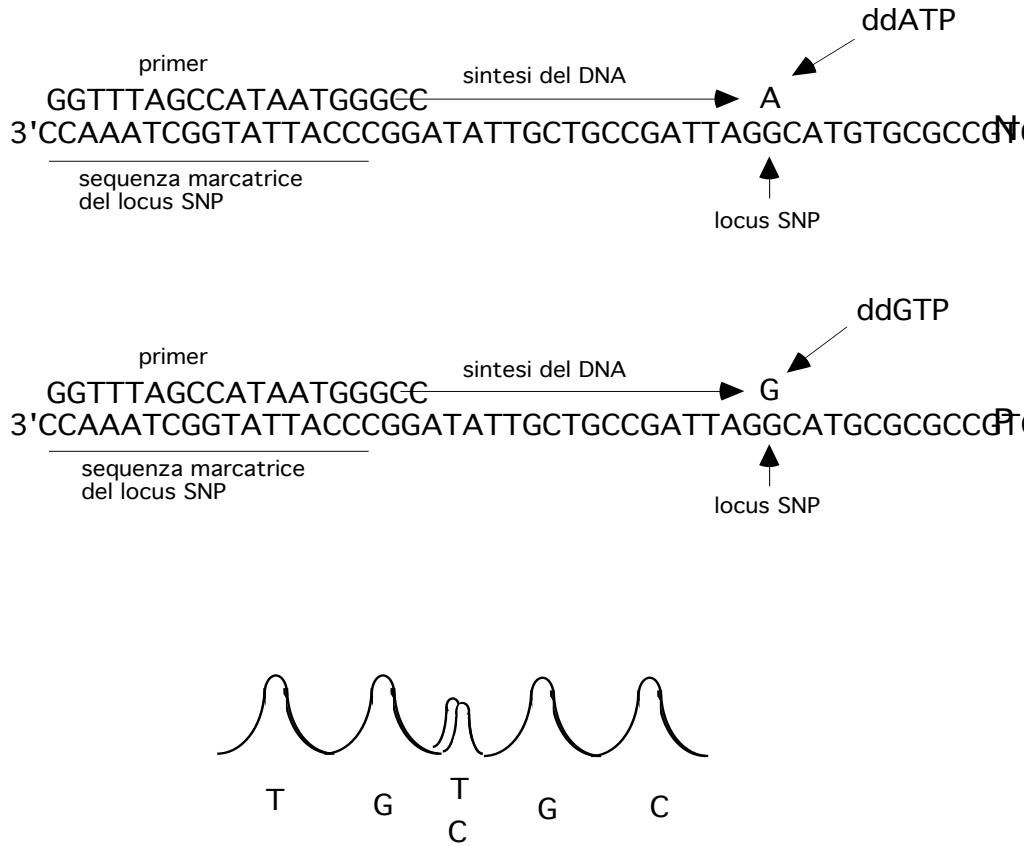



Figura 3-10. Identificazione del marcatore genetico SNP mediante determinazione della sequenza nucleotidica del DNA del locus SNP. Il DNA estratto da cellule diploidi viene sottoposto ad analisi della sequenza utilizzando primer complementari alla sequenza marcatrice del locus del polimorfismo SNP ed i 4 ddNTP fluorescenti. Le sequenze dei due alleli risulteranno identiche ad eccezione di un'unica posizione in cui sono presenti due basi diverse (T e C) indicate da due diversi ddNTP. In figura sono indicati solo ddGTP e ddTTP. N = allele normale, P = allele mutato patologico. Nella figura in basso è indicato il grafico dei picchi di cinque posizioni della sequenza con il locus SNP in posizione centrale. L'identità della basi è data dal diverso colore di fluorescenza dei ddNTP (non indicato in figura). Il grafico mostra che nella posizione centrale (sito SNP) sono presenti due picchi appartenenti a due basi diverse, T e C. Il loro picco è più basso degli altri perché la quantità di basi stampo per quelle due basi eterozigoti è la metà della quantità delle altre basi stampo.

Tabella. 3-2. Marcatori del genoma nucleare umano

Marcatori	Numero di loci	Numero di alleli	Distanza media dei loci
	Marcatori genetici		
Gruppi del Sangue[1]	1	3	---
Sequenze RFLP	>100.000	2	circa 10.000b (tutti i siti RFLP) circa 300.000b (un tipo di sito RFLP)
Sequenze VNTR minisatellite	>10.000	molti altamente informativi	circa 100.000b
Sequenze VNTR microsatellite	>10.000	molti altamente informativi	circa 30.000b
Base SNP	>1.000.000	4 spesso 2	circa 1.000b
	Marcatori fisici		
Sequenze STS	>30.000	nessuno-pochi[2]	100.000b[3]
Sequenze EST	>10.000	nessuno-pochi[2]	40.000b[3]
Ogni marcatore genetico sopra indicato diviene marcatore fisico quando il suo locus viene individuato in una mappa fisica			

1) Sono considerati solo gli alleli dei principali gruppi del sangue: A, B,ed 0.
2) EST e STS sono marcatori fisici non necessariamente polimorfici, tuttavia alcune di queste sequenze sono polimorfiche e possono essere usate come marcatori genetici. La caratteristica di marcatore genetico o fisico è data dall'unità di misura in cui è misurata la loro distanza dal marcatore più vicino al termine del braccio p del cromosoma. cM per i marcatori genetici e numero di basi per i marcatori fisici.
3)Per le sequenze EST e STS sono indicate le distanze medie che saranno raggiunte quando saranno tutte mappate.
(da Strachan T. and Read A.P. (1999) Human Molecular Genetics. 2a Ed., Bios, UK. ridisegnata e modificata).

SNP. La sequenza includente il locus SNP è determinata con il metodo ciclico automatizzato utilizzando ddNTP fluorescenti (capitolo 1) e la base SNP mutata è confrontata con quelle di alleli noti. In figura 3-10 è mostrato il grafico della sequenza ottenuta mediante determinazione ciclica utilizzando ddNTP fluorescenti (capitolo 1) ed in figura 1-10c è mostrata l'analisi di un SNP con il metodo manuale di Sanger.
Le analisi della sequenza del DNA fatte per analizzare i marcatori SNP, essendo eseguite sul DNA genomico estratto da cellule diploidi di un individuo, mostrano l'identità delle due varianti, la loro condizione di omozigosi o di eterozigosi ma non la loro origine materna o paterna. Essa può essere individuata solamente analizzando il DNA dello stesso locus SNP dei genitori o dei figli dell'individuo di interesse (figura 3-16).

Mappe genetiche del genoma umano

La scoperta dei marcatori genetici polimorfici ripetuti nel genoma, RFLP, VNTR e SNP ha permesso di costruire con ognuno di questi tipi di marcatori delle mappe genetiche di associazione (tabella 3-3). Le tre mappe sono state integrate elettronicamente per ottenere un'unica mappa genetica di ricombinazione altamente dettagliata. Una mappa genetica è la descrizione grafica della disposizione su ciascun cromosoma (disegnato come un segmento) dei loci dei marcatori genetici con scritta la distanza genetica in cM iniziando a contare dal marcatore genetico più vicino al termine del braccio piccolo (braccio p) del cromosoma (figura 3-14). Per semplificare la lettura, di lato alla mappa genetica di ogni cromosoma viene disegnata la riproduzione grafica dello stesso cromosoma metafasico con linee che nelle due mappe uniscono i due loci (citogenetico e genetico) di uno stesso gene o di uno stesso marcatore (figura 3-14).
Le mappe genetiche hanno semplificato molto la mappatura dei geni e delle patologie perché consultandole si ha la visione del locus e della sequenza dei marcatori di ogni cromosoma ed anche delle sequenze dei primer per individuare i marcatori mediante la tecnica PCR o mediante analisi ciclica della sequenza. Per mappare un gene o una patologia è sufficiente verificare la loro associazione con più marcatori genetici di ciascun cromosoma nei membri di una o più famiglie. Individuata l'associazione del gene ad un marcatore di un dato cromosoma, si individuano altri marcatori della mappa che abbiano con il gene di interesse valori di ricombinazione inferiori a quello del primo marcatore individuato, fino a costruire una mappa genetica dettagliata intorno al gene di interesse.
La prima mappa genetica umana costruita con sequenze polimorfiche fu proposta e costruita con marcatori RFLP da David Botstein nel 1980 (violinista alla Boston Simphony Orchestra e professore di genetica al Massachusetts Institute of Technology, Cambridge, USA) e da altri ricercatori.

Nel 1996 con il microsatellite (AC/GT)n, è stata costruita una mappa genetica di associazione avente 5.266 marcatori separati da una distanza media di 1,6cM. La regione di ogni VNTR AC/GT fu clonata da una genoteca genomica umana usando come sonda la sequenza ripetuta, per ogni clone furono determinate le sequenze al 5' ed al 3', e in ciascuna delle due sequenze fu individuata una sequenza di circa 20 basi unica nel genoma cioè specifica di quel dato locus.
Le due sequenze uniche furono utilizzate per disegnare i primer per la PCR analitica con la quale furono individuate le varie sequenze ripetute AC/GT poste su differenti loci (figura 3-9). Per questo furono analizzate genoteche genomiche costruite con il DNA di molti individui, appartenenti a famiglie diverse e si determinò la frequenza di ricombinazione dei nuovi marcatori con marcatori di locus noto e dei nuovi marcatori tra di loro, fino a stabilire per ciascun marcatore la distanza genetica dal telomero del braccio p di ciascun cromosoma. Con la stessa analisi fu possibile stabilire anche il grado di polimorfismo del marcatore di ciascun locus.
Questa mappatura del genoma umano è stata progressivamente migliorata con l'identificazione di nuovi loci di microsatelliti, i quali hanno una densità media calcolata di 1 locus ogni 30.000b.
Nel 1999 è stata iniziata la costruzione di una mappa di associazione utilizzando il polimorfismo SNP ed in 2 anni è stata messa a disposizione della comunità scientifica una mappa con 1.420.000 marcatori SNP, circa un SNP ogni 2kb. Il marcatore SNP, sebbene non sia altamente polimorfico (in genere ha solo due alleli) ha alcune caratteristiche che lo fanno preferire ai microsatelliti: i loci SNP possono essere analizzati con macchine automatiche (mentre è più difficile farlo per i microsatelliti), hanno alta densità nel DNA genomico (circa 1 SNP ogni 1kb contro le 30kb dei microsatelliti) e pertanto la mappa SNP potrà essere resa ancora più densa di marcatori. Al finanziamento della costruzione della mappa SNP hanno partecipato una decina di imprese private, molte delle quali farmaceutiche, perché con gli SNP si vogliono mappare i geni della suscettibilità a patologie complesse (capito 4) e fare studi di farmacogenetica.

<u>Associazione gametica preferenziale</u> (Linkage disequilibrium).
L'associazione gametica preferenziale è l'associazione di due alleli di due geni posti sullo stesso cromosoma che nella popolazione risultano associati più frequentemente di quello che ci si aspetterebbe se gli eventi di ricombinazione avvenissero a caso. Questa particolare associazione è spiegata assumendo che due geni A e B siano posti sullo stesso cromosoma ed un individuo AB/aB, subisca nelle sue cellule germinali una mutazione (evento fondatore) su un allele del gene B che causa la formazione di un nuovo allele non patologico (B*). Questo nuovo allele (B*) risulta associato all'allele A solo su quel cromosoma di quell'individuo che produrrà inizialmente gameti e quindi prole AB* e aB. Data la casualità degli eventi di ricombinazione durante la meiosi,

l'allele B* tenderà a distribuirsi in maniera uguale tra i due cromosomi omologhi formando un uguale numero di gameti AB* e aB*.
Se il portatore di questa nuova combinazione di alleli sarà il capostipite fondatore di una lunga (migliaia di anni) e grande (famiglie con molti figli) discendenza, nei suoi discendenti, meiosi dopo meiosi, ricombinazione dopo ricombinazione, fecondazione dopo fecondazione, generazione dopo generazione, la popolazione sarà costituita da un ugual numero di individui AB* ed individui aB*. Prima di allora l'allele B* sarà in associazione gametica preferenziale con l'allele A, cioè il numero dei portatori dell'allele AB* sarà superiore a quello dei portatori dell'allele aB*.
L'associazione gametica preferenziale dipende da due fattori: la distanza genetica tra i loci dei due geni e la distanza cronologica dalla mutazione cioè il tempo intercorso dal momento in cui è avvenuta la mutazione ed il momento nel quale si analizzano gli alleli. Le due distanze contribuiscono all'equilibrio degli alleli che è più difficile raggiungere quanto più le due distanze sono piccole. Se la mutazione è recente e gli alleli ricombinano raramente, occorreranno molte meiosi e quindi molto tempo prima che sia raggiunto l'equilibrio tra i due alleli.
E' stato osservato che un marcatore genetico è associato al 70% dei cromosomi portatori del gene patologico dell'anemia delle cellule falciformi (autosomica recessiva) e al 3% dei cromosomi sani. Le percentuali diverse indicano una associazione gametica preferenziale del marcatore con il gene mutato.

Mappa degli ibridi di radiazione.
Questo tipo di mappa è costruito utilizzando la tecnica di ibridi cellulari interspecifici (figura 3-4 e Tabella 3-3). Inizialmente, fine degli anni '60, ebbe un uso limitato perché erano noti pochi marcatori genetici. Successivamente dagli anni '90, con la clonazione di molti marcatori genetici ha avuto un grande sviluppo ed è stata molto utile per collegare sequenze di frammenti di DNA clonati provenienti da regioni di uno stesso cromosoma poste molto distanti tra loro (figura 3-12b).
Un procedimento per costruire una mappa di radiazione consiste nell'irradiare con raggi X gli ibridi cellulari di criceto contenenti un solo cromosoma umano. Le radiazioni causano la rottura in più frammenti del cromosoma umano e dei cromosomi di criceto. Le cellule irradiate sarebbero destinate a morire ma sono recuperate fondendole con altre cellule di criceto. Le cellule poi sono coltivate per aumentarne il numero e clonate in modo che ogni clone includa un solo frammento di cromosoma umano. Cloni diversi di uno stesso cromosoma umano includeranno frammenti diversi, ma in parte sovrapponibili, per cui sarà possibile ricostruire la loro disposizione originale nel cromosoma verificando la presenza di uno stesso marcatore genetico o fisico in due frammenti diversi.
Le distanze tra marcatori sono dette "distanze di radiazione" e sono determinate sulla base della frequenza delle rotture tra coppie di marcatori che è in relazione all'intensità della radiazione usata. Più i marcatori sono distanti

tra loro più è probabile che i marcatori si trovino su frammenti di radiazione diversi. Maggiore è la dose di radiazione usata, più piccoli sono i frammenti di cromosoma, ed i marcatori presenti in esso sono più vicini.
Il modo per costruire la mappa di radiazione è analogo a quello usato per costruire le mappe di associazione dove la frequenza dei crossing over, che sono anche essi rotture di cromosoma, è usata per misurare le distanze tra coppie di marcatori.
Nelle mappe di radiazione l'unità di misura è il centiRaggio (cR), analogo al centiMorgan, che è in relazione alla dose di radiazione usata per frammentare i cromosomi. Se per frammentare il cromosoma sono stati usati 8000rad (Radiation absorbed dose), l'unità di misura per quella mappa è il cR_{8000}.
La mappa di radiazione è stata completata ricostruendo ogni cromosoma con i frammenti nella posizione originale, sul cromosoma sono indicate la posizione dei marcatori e le distanze misurate in cR a partire dal marcatore più vicino al braccio piccolo (p) del cromosoma.
Per costruire la mappa di radiazione del cromosoma 21 umano che contiene 40Mb, il cromosoma è stato irradiato con 8000rad ed ha prodotto mediamente 5 frammenti di circa 5Mb. I frammenti avevano punti di rottura diversi che hanno permesso di ricostruire l'ordine dei frammenti nel cromosoma.
I frammenti di DNA ottenuti per radiazione includono lunghe regioni di DNA e permettono di collegare tra loro marcatori e geni mappati con altre strategie. Il confronto delle mappe di radiazione con quelle fisiche ha mostrato che esse indicano lo stesso ordine dei geni.
La mappa di radiazione è una mappa fisica, essa é basata sulla dimensione dei frammenti di radiazione. I frammenti hanno dimensione di molti Mb e questa dimensione rappresenta la risoluzione della mappa cioè la distanza media minima tra marcatori (tabella 3-3).

Costruzione delle mappe fisiche del genoma umano con distanze misurate in numero di basi

<u>La mappa dei contig dei cloni YAC.</u>
La mappa dei contig dei cloni YAC è stata costruita da Daniel Cohen e collaboratori nel 1993 ed aggiornata nel 1995. Essa è la prima mappa fisica umana con le distanze tra i marcatori misurate in numero di basi (3-14).
La mappa dei contig dei cloni YAC è la descrizione grafica, su ciascuno dei 24 cromosomi umani (disegnati come sottili rettangoli) dei frammenti di DNA genomico, dei quali è stata stabilita in numero di basi la dimensione e la distanza dal telomero del braccio p del cromosoma (locus fisico).
Al fine di facilitare la lettura, la mappa fisica di ogni cromosoma é disegnata vicina alla relativa mappa citogenetica, con linee che nelle due mappe uniscono i rispettivi loci (citogenetico e fisico) di uno stesso gene o di uno stesso marcatore (figura 3-14). Tutti i dati della mappa sono conservati in una banca consultabile via internet.

La mappa dei contig dei cloni YAC è stata costruita utilizzando una genoteca genomica in cellule di lievito. I cloni di questa genoteca genomica, includevano inserti di DNA nucleare umano e di questi è stata definita la posizione subcromosomica fisica (fig 3-11a, 3-12a-b). Vagliando la genoteca, usando una PCR analitica specifica per una qualsiasi sequenza del genoma umano (un gene, un cDNA, un marcatore genetico) dal clone isolato si ha l'indicazione della posizione subcromosomica della sequenza di interesse (figura 3-11b).
Data la grande dimensione del DNA nucleare umano, fu necessario costruire un vettore che potesse accettare frammenti di DNA maggiori di 200kb con lo scopo di costruire con essi una genoteca genomica di circa 30.000 cloni che dovevano contenere frammenti di DNA umano provenienti da 10 genomi aploidi. Come vettore fu costruito il cromosoma artificiale di lievito, cromosoma YAC (Yeast Artificial Chromosome) che ha del lievito i telomeri ed i centromeri ed una sequenza autonoma di replicazione di DNA di lievito (ARS); inoltre nel DNA cromosomico dello YAC furono inseriti siti di restrizione per poter inserire i frammenti di DNA umano e marcatori per poter identificare il DNA degli YAC da quello dei cromosomi naturali delle cellule di lievito. Il DNA genomico fu digerito parzialmente con un enzima avente relativamente pochi siti di taglio sul DNA nucleare umano ed i frammenti con meno di 200kb scartati. I frammenti di DNA umano inseriti negli YAC avevano una dimensione media di 1Mb. La parziale digestione del DNA produsse frammenti di DNA umano parzialmente sovrapponibili (figura 3-12) al fine di poter ricostruire con essi il DNA del cromosoma (figura 3-12a-b). Gli YAC transfettati in cellule di lievito replicano in maniera lineare come i cromosomi della cellula ospite ma più frequentemente.
La genoteca genomica con cui è stata costruita la prima mappa fisica includeva 33.000 cloni YAC e l'inserto di DNA umano di ogni clone YAC fu sequenziato completamente o solo agli estremi 5' e 3'.
La frammentazione del DNA genomico umano per costruire la genoteca YAC aveva fatto perdere le posizioni che i frammenti avevano nelle molecole di DNA dei cromosomi. I costruttori della genoteca YAC, pazientemente ed abilmente, individuarono la posizione subcromosomica originale degli inserti di DNA umano presenti nei 33.000 cloni riuscendo a mappare il 75% del DNA genomico umano. Le regioni di DNA cromosomico furono ricostruite utilizzando le porzioni sovrapponibili degli inserti dei cloni YAC individuate mediante ibridazione molecolare o *in silice*. Queste regioni di DNA ricostruite che avevano una dimensione media 10Mb furono chiamate **contig** (figure 3-12 e 3-14).
Fu calcolato che complessivamente tra tutti i contig di tutti i cromosomi le regioni non ricostruite erano tra 200 e 1000, pertanto con i contig non si potevano ricostruire ampie regioni cromosomiche, tanto meno interi cromosomi. Da ciò la necessità di utilizzare altre strategie per ricostruire le regioni di DNA cromosomico mancanti tra contig diversi.
Una strategia adottata consisteva nell'utilizzare i marcatori di mappe genetiche per tipizzare cloni e contig. I marcatori genetici sono punti fissi nella sequenza del DNA dei cromosomi e, individuandoli in un inserto di YAC o in un contig, permettevano di caratterizzarli (tipizzazione), di individuare la loro posizione

sub-cromosomica e di assemblare inserti e contig che avevano un marcatore genetico in comune.
Una volta inseriti in contig della mappa fisica i marcatori genetici assumevano le caratteristiche di marcatori fisici e ad essi fu attribuita la sigla di STS (Sequence Tagged Site, sito marcato da sequenza).
I marcatori STS identificati da Cohen erano polimorfici perché originati da marcatori genetici, successivamente altri costruttori di mappe fisiche individuarono e caratterizzarono molti marcatori STS non polimorfici (vedere dopo). L'esistenza in più forme alleliche non è una caratteristica necessaria per un marcatore fisico perché la sua distanza da un altro marcatore fisico è in numero di basi (e non per frequenza di ricombinazione).
I marcatori STS nella formulazione attuale hanno le seguenti caratteristiche:
sono sequenze uniche nel genoma, di piccole dimensioni (meno di 500b), hanno un definito orientamento nel DNA dei cromosomi (il loro 5' è orientato verso il telomero del braccio piccolo del cromosoma), una definita posizione subcromosomica (distanza in numero di basi dal telomero del braccio p dei cromosomi) ed a ciascuno dei loro estremi 5' e 3' è stata definita una sequenza unica nel genoma di circa 20 basi, al fine di potere individuare il marcatore STS rapidamente ed inequivocabilmente mediante PCR analitica, cioè mediante analisi elettroforetica dell'amplificato PCR.
Un'altra strategia usata dal gruppo di Cohen fu quella di utilizzare come sonde 500 inserti di cloni YAC ciascuno dei quali includeva marcatori STS polimorfici, al fine di mapparli citogeneticamente mediante FISH. In questo modo furono coordinate le tre mappe del genoma umano: fisica, genetica e citogenetica.
La mappa fisica costruita era un'opera di ingegneria genetica di alto livello, tuttavia aveva dei limiti: ogni cromosoma aveva, tra i contig, parti non ricostruite che totalmente ammontavano al 25% del genoma; inoltre il 40-50% dei cloni YAC avevano un inserto chimerico. Gli inserti chimerici sono formati da frammenti di DNA provenienti da cromosomi diversi, che nel costruire la genoteca genomica, accidentalmente sono legati covalentemente per uno dei loro estremi e quindi inseriti nello stesso YAC. Il chimerismo degli inserti causa complicazioni nella mappatura fisica dei geni (figura 3-14c).
Tutti i dati genetici e molecolari dei cloni YAC (cromosoma di appartenenza, dimensione della sequenza, presenza di marcatori) furono catalogati, depositati in una banca dati elettronica e messi gratuitamente a disposizione della comunità scientifica mondiale. La genoteca genomica fu distribuita a varie istituzioni scientifiche. In Italia all'Università di Pavia e al DIBIT, S.Raffaele di Milano, che provvedevano gratuitamente a mappare geni e cDNA di chi ne faceva richiesta.
La costruzione della genoteca YAC e la catalogazione ordinata dei loci dei suoi oltre 33.000 cloni è stata una grande conquista del 20° secolo, essa ha aperto la via alla costruzione di altre mappe fisiche ed alla progettazione e realizzazione del progetto genoma umano: la determinazione della sequenza del DNA di tutti i cromosomi umani.

Mappatura fisica dei geni umani mediante vaglio della genoteca YAC.

La genoteca YAC è vagliata mediante PCR analitica specifica per il gene di interesse. Sono usati primer complementari a regioni del gene che sono uniche nel genoma. L'isolamento del relativo clone porta ad individuare il locus del gene di interesse.

In figura 3-11a è mostrata una genoteca YAC costituita da 23040 cloni. Al fine di ridurre il numero di analisi, le colonie dei vari cloni sono state riunite in 60 provette. Il contenuto di ciascuna provetta, chiamato "insieme di colonie" include la sospensione delle cellule di colonie originate da 394 differenti cloni YAC. Ordinatamente e progressivamente il contenuto di tre delle 60 provette è stato riunito in una nuova provetta, ottenendo così un totale di 20 provette. Il contenuto di ognuna delle 20 provette, che include le cellule delle colonie originate da 1152 differenti cloni YAC, è stato chiamato "insieme di insiemi di colonie". In questo modo il vaglio iniziale della genoteca YAC è ridotto da 23040 a 20 analisi di PCR analitica. La PCR analitica è specifica per il gene di interesse perché la sequenza dei primer è stata disegnata utilizzando la sequenza del gene di interesse o del suo cDNA.

In figura 3-11b è mostrato il vaglio della genoteca YAC che inizia con l'analisi PCR dei 20 "insieme di insiemi di colonie". L'analisi PCR è fatta direttamene sulle cellule delle colonie di lievito, perché la temperatura di 95°C distrugge le strutture cellulari che liberano il DNA dei cromosomi permettendo l'azione della miscela di reazione della PCR.

L'analisi dell'amplificato mediante elettroforesi su gel di agarosio indica quale dei 20 "insiemi di insiemi di colonie" YAC è positivo (cioè include la sequenza umana di interesse, 1-3 in figura 3-11b). Avuta questa informazione si ripete con gli stessi primer la PCR analitica sul DNA dei 3 singoli "insiemi di colonie" (1, 2 e 3 in figura 3-11b) che erano riuniti nello "insieme di insiemi di colonie" risultato positivo. Con questa nuova analisi è individuato quello dei tre "insieme di colonie" di cloni YAC che contiene il DNA di interesse (1 in figura 3-11b).

Le colonie originate degli stessi 394 cloni YAC dello "insieme di colonie" precedentemente erano state trasferire dal terreno solido di coltura su filtri di nitrocellulosa. Le 394 colonie erano state disposte su 4 filtri, chiamati piastre, ed in ogni piastra le colonie erano disposte ordinatamente come su una scacchiera.

Individuato l'"insieme di colonie" positivo si utilizza il suo DNA come stampo

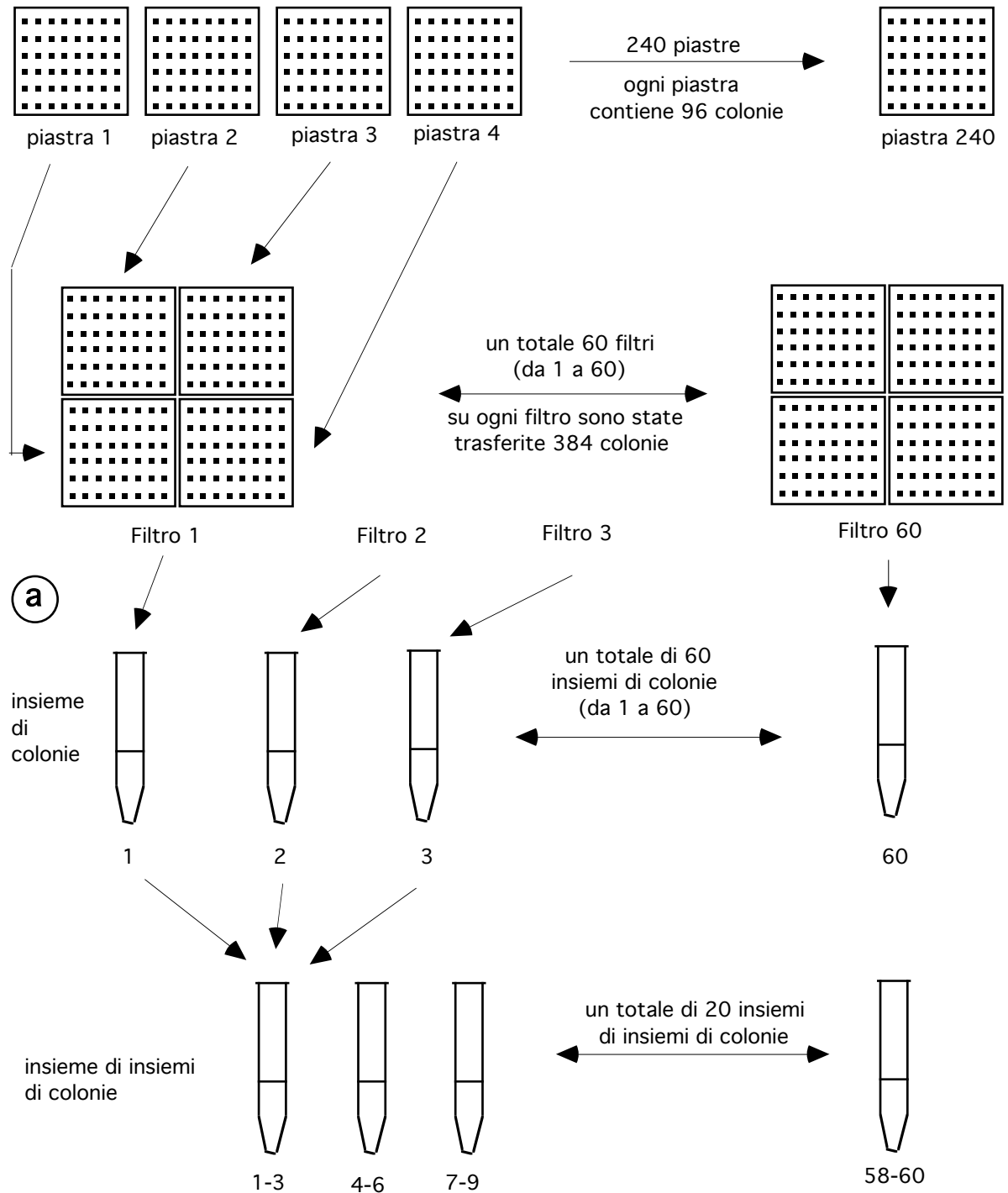



Figura 3-11. Preparazione dei campioni e vaglio della genoteca YAC.
a) preparazione dei campioni per il vaglio della genoteca YAC.
b) vaglio della genoteca YAC (pagina seguente).
L = DNA totale di lievito; U = DNA totale di uomo. Per altri dati vedere il testo.
(ridisegnato e modificato da Watson J.D., Gilman M., Witkowski J. and Zoller M. (1992) Recombinant DNA, 2nd ed., Scientific American Books, Freeman, USA).

insieme di insiemi di colonie
un totale di 20 insiemi di insiemi di colonie
1-3
4-6
7-9
58-60
Analisi del DNA di ciascuno dei 20 insieme di insiemi di colonie mediante PCR analitica specifica per il gene di interesse
PCR su ciascuno dei 20 insiemi di insiemi di colonie
1-3 4-6 7-9
58-60 L U
Elettroforesi degli amplificati PCR
L = DNA di lievito
U = DNA di uomo
positivo
b
Ripetizione della PCR analitica specifica per il gene di interesse sul DNA singoli insiemi di colonie 1, 2 e 3
1-3 1 2 3 L U
Elettroforesi degli amplificati PCR
positivo
Ibridazione dell'amplificato PCR del DNA dell'insieme 1 con il filtro 1, che contiene in spot distinti il DNA delle colonie originate dai 394 cloni che costituiscno l'insieme di colonie 1
autoradiografia
piastra 2
piastra 3
piastra 1
piastra 4
filtro 1
possibile recupero dalla piastra del DNA del clone YAC positivo
piastra 4

per amplificare mediante PCR analitica specifica per il DNA del gene di interesse. Il DNA amplificato poi è utilizzato come sonda per vagliare mediante ibridazione il DNA disposto ordinatamente su filtro di ciascuno dei 394 cloni YAC che formano lo "insieme di colonie" risultato positivo. Un solo clone deve risultare positivo e dalla sua posizione sul filtro si ha l'indicazione del locus fisico del gene di interesse, dato che di ogni clone YAC sono noti la posizione sul filtro e la posizione subcromosomica dell'inserto di DNA umano dello stesso clone YAC.
Il DNA del clone YAC positivo può essere recuperato dalla piastra per fare su di esso ulteriori verifiche di identità.

La mappa delle sequenze STS.
Il grande lavoro di Cohen e colleghi permise ad altri ricercatori di costruire un'altra mappa fisica più densa di marcatori STS. La quasi totalità degli STS (siti marcati da sequenze) fu individuata e caratterizzata dai costruttori della nuova mappa fisica.
La mappa delle sequenze STS è la descrizione grafica dei loci (loci fisici) dei marcatori STS con le distanze dei loci misurate in numero di basi dal telomero del braccio p di ogni cromosoma (disegnato come un segmento). Tutti i dati della mappa sono stati conservati in una banca dati al fine di poterli consultare, di poter ricercare i marcatori STS mediante programmi che simulano l'ibridazione (ibridazione elettronica) oppure di conoscere le sequenze locus specifiche di un dato marcatore STS per ricercarlo in una genoteca genomica mediante PCR analitica.
I ricercatori per ricostruire l'assetto originale dei frammenti di DNA genomico inseriti nei dei vari cloni YAC utilizzarono i dati della prima mappa fisica; inoltre vagliando la genoteca YAC di Cohen individuarono 10.850 nuovi marcatori STS, la maggioranza dei quali non era polimorfico, e costruirono una mappa di radiazione contenente 6193 marcatori citogenetici che integrarono con una mappa genetica contenente 5264 marcatori genetici. Utilizzando la mappa di radiazione integrata con i marcatori genetici fu possibile costruire dei contig sempre più lunghi che coprivano il 95% del DNA dei cromosomi, fino a definire la mappa fisica con 15.000 marcatori STS tra loro distanti mediamente 200kb (figure 3-12a e b, 3-14 e tabella 3-3).
L'alta densità di marcatori STS permetteva di mappare più precisamente i geni (come prendere le distanze tra due punti con un metro diviso in centimetri o con un altro diviso solo in decimetri). Inoltre l'alta densità di marcatori STS permise di individuare i cloni YAC che avevano inserti chimerici e di individuare la posizione subcromosomica dei due frammenti di DNA che costituivano il frammento chimerico. Successivamente per stabilire la corretta sequenza del DNA delle regioni cromosomiche da cui provenivano i due frammenti di DNA furono vagliate altre genoteche genomiche usando PCR analitiche specifiche per le sequenze STS presenti negli stessi frammenti. Per questi scopi furono

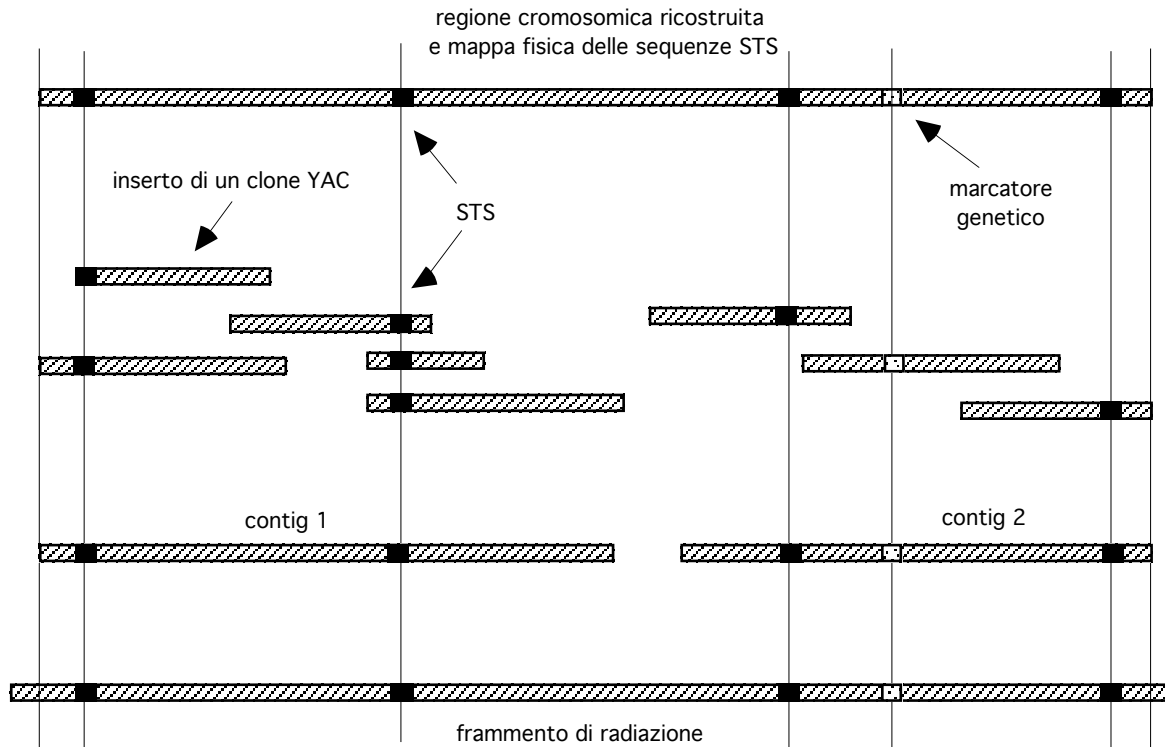



Figura 3-12a. Schema riassuntivo dei metodi usati per la ricostruzione di una regione cromosomica di una mappa fisica.
Mediante sovrapposizione parziale degli inserti di DNA genomico umano dei cloni YAC e la presenza in essi di marcatori STS sono stati costruiti due contig. L'assenza di contiguità tra i due contig è superata allineando le sequenze STS dei due contig con quelle identiche presenti nel frammento di radiazione. Le sequenze STS, essendo sequenze uniche nel genoma, permettono di stabilire con sicurezza le posizioni relative dei frammenti di DNA e quindi anche dei due contig. Il frammento di radiazione, del quale è nota l'origine cromosomica, conferma la localizzazione cromosomica dei contig e permette di ricostruire la sequenza del DNA cromosomico della regione nel caso che un frammento di DNA sia stato perso durante la costruzione della genoteca YAC.
I criteri dei metodi usati per la costruzione della mappa fisica dei contig dei cloni YAC e quello per la costruzione della mappa fisica delle sequenze STS sono simili.
La mappa fisica dei contig di Cohen fu costruita prevalentemente mediante sovrapposizione di sequenze appartenenti ad inserti diversi di cloni YAC aventi sequenze identiche ad uno dei loro estremi, in parte mediante l'uso di marcatori STS polimorfici comuni a due frammenti diversi di cloni YAC e mediante ibridazioni *in situ* con FISH.
La mappa fisica delle sequenze STS fu costruita utilizzando le mappature fisiche della prima mappa dei contig di Cohen, la definizione di marcatori genetici dei quali era stato individuato il locus fisico mediante integrazione della mappa genetica con una di radiazione e la definizione di molti nuovi marcatori STS non necessariamente polimorfici. Questa metodologia portò alla costruzione di contig che coprivano la quasi totalità del DNA dei cromosomi. Per altri dettagli vedere il testo.

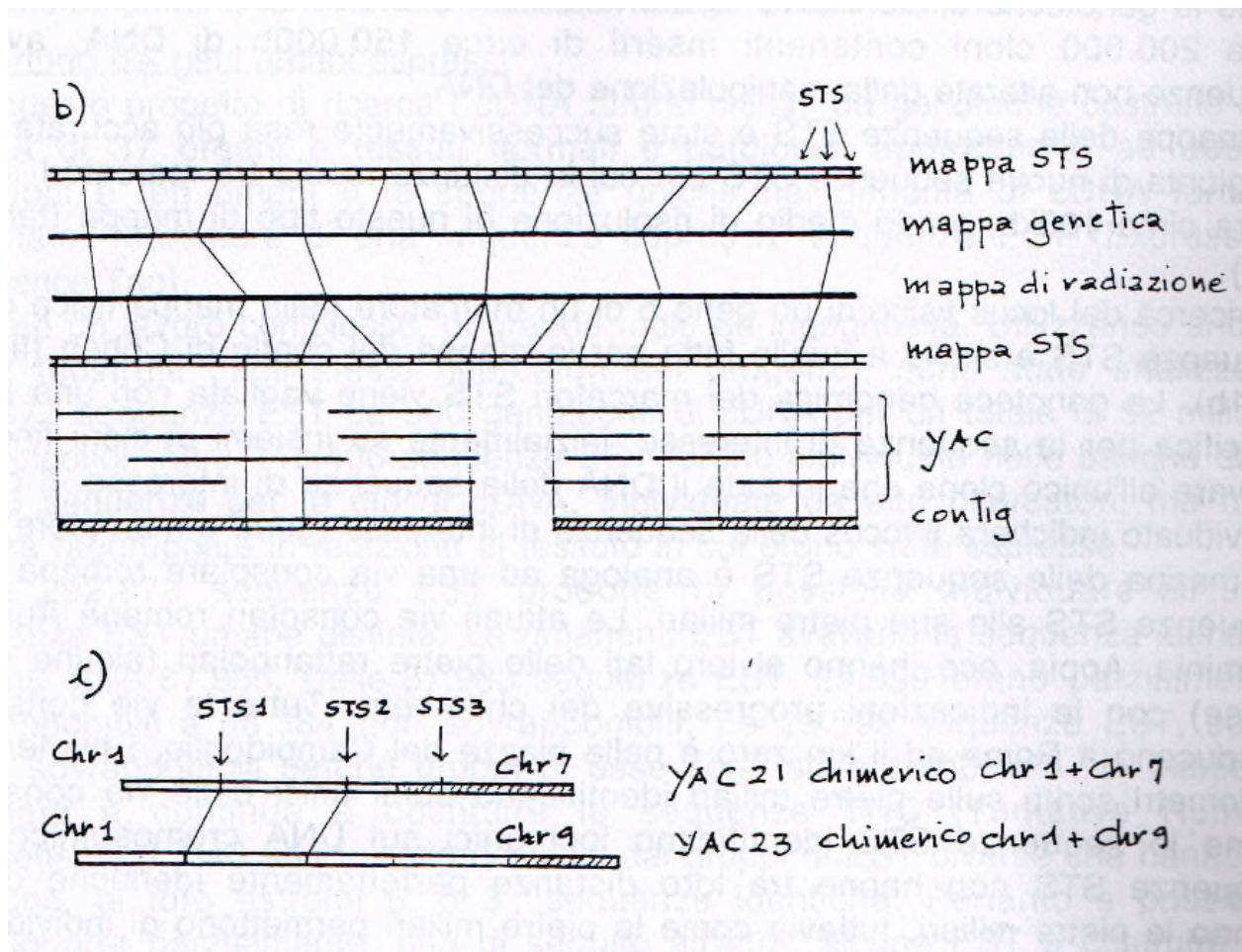



Figura 3-12 b). Strategia per la costruzione della mappa delle sequenze STS.
La mappa delle sequenze STS è stata integrata con la mappa genetica e quella di radiazione. In figura i loci nelle tre mappe dello stesso marcatore STS sono collegati da linee sottili. L'integrazione della mappa genetica con la mappa fisica permette di stabilire il locus fisico dei marcatori genetici (marcatori STS polimorfici). L'integrazione della mappa di radiazione, che ha frammenti cromosomici molto lunghi, con la mappa fisica permette di collegare contig tra i quali manchi un tratto di DNA cromosomico. L'interruzione tra i due contig rappresenta una regione cromosomica persa durante la manipolazione del DNA. Per comodità di confronto la stessa mappa delle sequenze STS è stata disegnata due volte.
c) Quando su un YAC sono stati individuati solo due diversi marcatori STS (single linked STS) il tratto di DNA dell'inserto YAC non è considerato sicuro perché a causa del chimerismo degli YAC un marcatore STS potrebbe appartenere al DNA di un cromosoma e l'altro marcatore STS ad un altro cromosoma (es. in c, STS2 ed STS3 sono sullo stesso inserto YAC ed appartengono a cromosomi diversi). Mentre se due diversi marcatori STS sono presenti su almeno 2 inserti di cloni YAC diversi (double linked STS) il collegamento è considerato sicuro perché è molto improbabile che gli inserti di due YAC diversi siano chimerici per gli stessi segmenti di DNA (in c, STS1 ed STS2 sono ambedue presenti sugli YAC 21 e 23). Il chimerismo dipende da una legatura casuale di due pezzi di DNA appartenenti a cromosomi diversi o a zone diverse di uno stesso cromosoma. La legatura avviene durante la manipolazione del DNA genomico per costruire la genoteca YAC. (da Hudson T.J. et al. Science 270, 1945-1954, 1995. ridisegnato e modificato).

vagliate genoteche genomiche di una nuova generazione aventi lunghi inserti come le genoteche di cloni BAC (Bacterial Artificial Chromosome) costituite da circa 200.000 cloni contenenti inserti di circa 150.000b di DNA, aventi sequenze non alterate dalla manipolazione del DNA.
La mappa delle sequenze STS è stata successivamente resa più accurata con l'aggiunta di nuove sequenze STS per cui la distanza media tra marcatori STS è ora circa 100kb, grado medio di risoluzione di questo tipo di mappa (tabella 3-3).
La ricerca del locus fisico di un gene o di un marcatore nella mappa fisica delle sequenze STS è simile a quella fatta per la mappa dei contig di Cohen (figura 3-11b). La genoteca genomica dei marcatori STS viene vagliata con una PCR specifica per la sequenza di interesse, inizialmente su insiemi di cloni fino ad arrivare all'unico clone che include il DNA della sequenza di interesse. Il clone individuato indicherà il locus della sequenza di interesse (gene o marcatore).
La mappa delle sequenze STS è analoga ad una via consolare romana e le sequenze STS alle sue pietre miliari. Le attuali vie consolari romane Aurelia, Flaminia, Appia, ecc. hanno ai loro lati delle pietre rettangolari (alcune sono perse) con le indicazioni progressive dei chilometri. Tutte le vie consolari conducono a Roma ed il km zero è nella piazza del Campidoglio. I numeri dei chilometri scritti sulle pietre miliari identificano punti unici delle vie consolari come le sequenze STS identificano loci unici sul DNA cromosomico. Le sequenze STS non hanno tra loro distanze perfettamente identiche come hanno le pietre miliari, tuttavia come le pietre miliari permettono di individuare un unico punto (locus) nella lunga molecola del DNA dei cromosomi.

Mappa di restrizione del genoma umano.
Sono state costruite delle mappe fisiche utilizzando come marcatori i siti di taglio di enzimi di restrizione la cui sequenze palindromiche erano relativamente scarse nel genoma umano. La dimensione dei frammenti ottenuti dopo digestione è di circa 200-400kb. Pertanto questa dimensione è il grado di risoluzione in numero di basi di questo tipo di mappa fisica (tabella 3-3)

Mappa fisica delle sequenze espresse.
Nel 1995, con il progetto genoma umano ancora in corso, un gruppo di 78 ricercatori clonò, sequenziò e catalogò 87.983 sequenze di cDNA uniche nel genoma, per lo più parziali (non codificanti l'intero mRNA) provenienti da molti tessuti umani.
Lo scopo della loro costruzione era di analizzare solo la parte codificante dei geni al fine di arrivare a conoscere più rapidamente il numero dei geni umani, le loro sequenze codificanti e le sequenze delle proteine sintetizzate nei tessuti umani normali e patologici. In questo modo si poteva procedere più rapidamente alla definizione di tutti i geni umani perché si evitava di analizzare il DNA non codificante che costituisce più del 90% del DNA genomico. La ricerca portò ad avere un elenco della quasi totalità delle sequenze espresse dai geni umani che fu chiamato repertorio dei geni umani espressi.

Repertorio dei geni umani espressi.

Per questo progetto di ricerca i cDNA furono clonati da genoteche costruite da mRNA di 37 organi e tessuti normali e patologici appartenenti ad adulti, embrioni e feti umani e le sequenze uniche nel genoma di cDNA furono chiamate "marcatore di una sequenza espressa" sequenze EST (Expressed Sequence Tag).

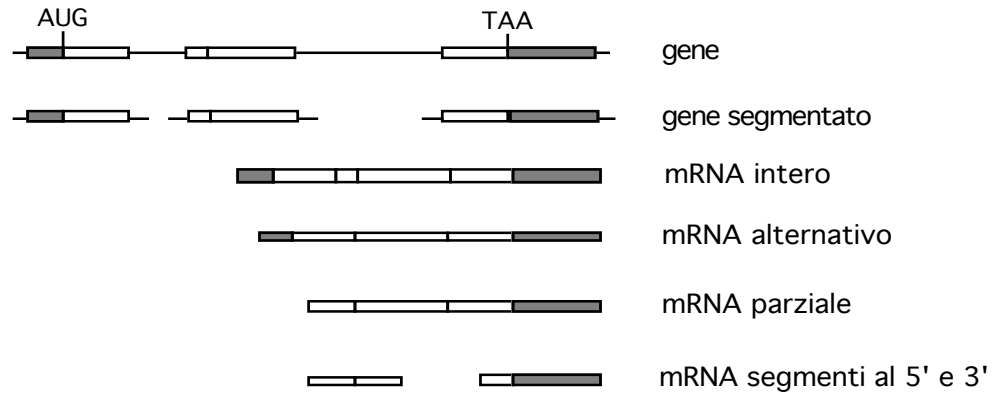
E' stato un lavoro complesso, di grande mole ed ingegnosità. Utilizzando anche robot (costruiti dagli stessi ricercatori del progetto), sono state analizzate 174.472 sequenze EST da 300 genoteche di cDNA per un totale di 52 milioni di nucleotidi. Altre 118.406 sequenze EST furono individuate nelle banche dati, erano sequenze per lo più di cDNA, individuate da altri ricercatori, ma non ancora raggruppate in relazione al tessuto in cui erano state espresse.

Delle 292.878 sequenze EST prodotte fu possibile individuare 87.983 sequenze EST umane distinte. Le rimanenti EST avevano la sequenza identica ad una delle 87.983. Delle 87.983 sequenze EST, 29.599 erano parzialmente sovrapponibili e 58.384 non sovrapponibili. Le 29.599 sequenze EST, sono dette sovrapponibili perché gruppi di esse sono state combinate in maniera ordinata e continua a costituire le sequenze THC (Tentative Human Consensus). I THC sono contig costituiti dai gruppi di EST diverse che hanno in comune, ai loro estremi 5' o 3', sequenze identiche. Pertanto è possibile sovrapporre parzialmente, una di seguito all'altra, alcune EST e costruire un THC ed il THC è completo quando è ricostruito il cDNA di un intero mRNA. I THC furono costruiti utilizzando anche sequenze di cDNA provenienti da tessuti diversi, assumendo che gli mRNA da cui i cDNA derivavano, fossero il prodotto di uno stesso gene perché avevano identica una parte della loro sequenza (almeno 20 nucleotidi).

FUNZIONE dei GENI ESPRESSI[1]	%
Metabolismo: energetico, delle piccole molecole, dei cofattori.	16
Espressione genica: sintesi, degradazione e modificazioni covalenti del RNA e proteine.	22
Struttura e motilità cellulare: citoscheletro, microtubuli, matrice extracellulare.	8
Omeostasi e difesa cellulare e/o dell'organismo: immunologia, riparazione del DNA, risposta allo stress.	12
Meccanismi di comunicazione intercellulare: recettori, ormoni/fattori di crescita, trasduttori di segnale, effettori/modulatori di segnale.	12
Sintesi del DNA e divisione cellulare: sintesi del DNA, ciclo cellulare, apoptosi e struttura dei cromosomi.	4
Geni non classificati	25

Le 87.983 sequenze EST hanno portato ad identificare 10.124 geni già noti e le rimanenti 48.170 nuovi geni dei quali 15.475 come THC.
87.983 è un numero più alto della stima attuale del numero totale dei geni (circa 30.000), ma è possibile spiegarlo in due modi: più di una sequenza EST unica non sovrapponibile può essere originata da uno stesso mRNA (vedere figura 3-13b) o da uno stesso gene per splicing alternativo.
30 genoteche (tra le 37 analizzate) erano vagliate più dettagliatamente, usando più di 1000 sequenze EST. Sebbene questa analisi non includa tutti i geni umani, da essa si sono avute importanti informazioni. Essa ha mostrato che solo 8 geni sono espressi in tutti i 30 tessuti/organi e 227 geni erano espressi in 20-27 tessuti/organi.
Questi dati indicano un'alta ed imprevista specificità di espressione di geni in cellule diverse. Infatti solo una piccola frazione di geni umani (227 su circa 30.000) sono espressi nella quasi totalità dei tessuti umani. Si assume che questi geni siano coinvolti nel mantenimento in vita (house keeping) di ogni tipo di cellula e siano espressi in quantità abbondante o moderatamente abbondante.
Altri geni house keeping probabilmente sono espressi in quantità limitate o sono espressi solo in particolari stati fisiologici (es. digiuno, rialimentazione) per cui la scarsa concentrazione nelle cellule dei loro mRNA non ha permesso la sintesi dei relativi cDNA.

Mappa delle sequenze EST umane
Quando le mappe dei cloni YAC e delle sequenze STS furono messe a disposizione della comunità scientifica internazionale, molti ricercatori iniziarono a ricercare in esse i loci dei geni e dei cDNA da essi clonati. Anche le sequenze EST furono utilizzate per questo scopo ed hanno contribuito a migliorare la definizione delle stesse mappe fisiche.
Una mappa fisica delle sequenze EST è già stata pubblicata e conservata nelle banche dati. Come nelle altre mappe fisiche, la distanza delle sequenze marcatrici espresse è data in numero di basi iniziando a contare dal telomero del braccio p di ogni cromosoma. Essendo una mappa di sequenze espresse, esse saranno collocate esclusivamente all'interno dei geni: negli esoni e nelle parti trascritte ma non tradotte degli mRNA. Utilizzando la mappa delle EST la mappatura fisica di un gene, del quale si conosca la sequenza anche parziale o la sequenza del suo cDNA, è molto semplice da ottenere. E' sufficiente allineare elettronicamente la sequenza di interesse con quelle della mappa delle EST conservate nella banche dati ed il programma individuerà la EST identica alla sequenza di interesse. Quella EST è il marcatore fisico del gene di interesse.
La prima mappa fisica dei geni espressi risultava parziale a causa di alcuni problemi: perdita durante la costruzione delle genoteche di mRNA scarsamente espressi o espressi in momenti fisiologici particolari o comunque in momenti diversi (es. notte) da quelli del prelievo dei tessuti dai donatori.
Al fine di arrivare rapidamente ad una migliore definizione della mappa fisica dei geni espressi, un gruppo di ricercatori analizzò le regioni trascritte, ma non

a) Le diverse sequenze di uno stesso gene e del suo mRNA

b) STS ed EST di uno stesso gene

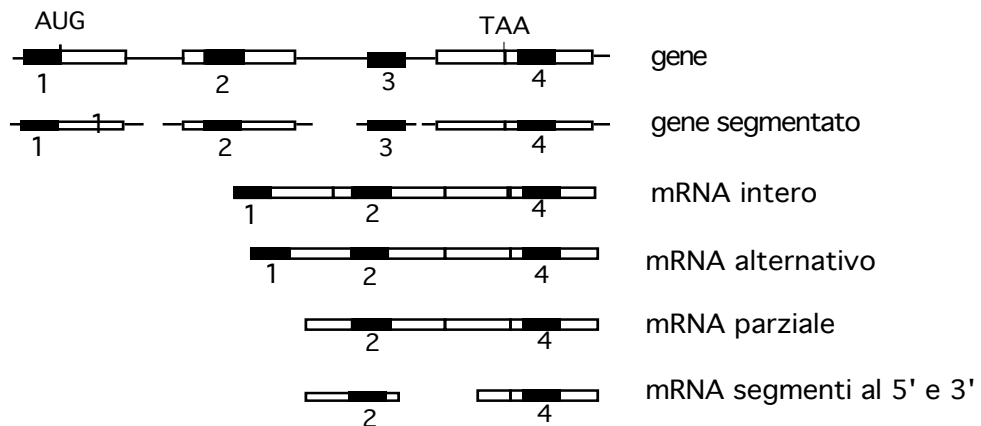



Figura 3-13. Lo schema mostra come dal DNA di un gene per reazioni naturali o manipolazioni genetiche si possano produrre frammenti di DNA genomico o del suo mRNA aventi sequenze totalmente diverse.
I segmenti di gene e di mRNA sono spezzoni che si possono formare durante i processi di estrazione e manipolazione genetica.
a) Le regioni del gene trascritte ma non tradotte sono ombreggiate.
b) Le sequenze STS sono indicate con rettangoli neri e numerate. STS-1, STS-2 e STS-4 sono anche EST. Le sequenze EST che sono anch'esse STS, sono marcatori genetici importanti perché oltre ad individuare un gene espresso individuano anche il locus. STS-3 è localizzata in un introne, pertanto non compare nel mRNA e può comparire all'interno di altri geni. La EST al 3' della regione non tradotta degli mRNA (n. 4 in figura) è una 3'UTR = 3' Untranslated Region. (ridisegnato e modificato da M.S. Boguski and G.D. Shuler, ESTablishing a human transcript map Nature Genetics, 10, 369-371, 1995).

tradotte al 3', del gene strutturale (3'UTR) di 5.734 geni umani già noti e conservati nelle banche dati ed individuò 3.125 3'UTR uniche nel genoma. L'insieme di queste particolari EST fu chiamato "Unigene set" (serie di Unigene)(figura 3-13).
La decisione di analizzare le regioni 3'UTR era basata sul fatto che le sequenze 3'UTR sono più specifiche per un dato gene perché più libere di evolvere rispetto a quelle codificanti. La casualità delle mutazioni ha reso le sequenze 3'UTR dei vari geni diverse tra loro e quindi, in genere ma non sempre, specifiche del gene a cui sono contigue al 3'. Le sequenze codificanti dei geni hanno la restrizione nell'uso delle basi imposta dal dover formare triplette per codificare proteine e su di esse il controllo della selezione naturale e questo crea ridondanze nelle sequenze codificanti. All'interno di una famiglia di geni è più probabile trovare differenze di sequenza al 3'UTR che nella sequenza degli esoni. Le sequenze al 5'UTR dei geni hanno maggiore probabilità di essere simili perché includono sequenze comuni a più geni come le sequenze della regione dei promotori e dell'inizio della trascrizione. Pertanto il DNA delle sequenze 3'UTR è utilizzato per individuare i geni perché è la regione del gene che ha la più alta probabilità di essere unica nel genoma e quindi specifica di quel gene.
Alcune delle sequenze EST non sovrapponibili con altre, quindi apparentemente appartenenti a geni diversi, risultarono appartenenti allo stesso gene (figura 3-13b) quando fu trovato che mappavano sullo stesso locus, cioè sul locus fisico del gene che le aveva espresse. In questo modo furono costituiti 14.457 raggruppamenti (cluster) di EST, originati da 33.675 EST.
I vari raggruppamenti furono chiamati serie UniEST che si assumeva corrispondessero a geni. Utilizzando questa strategia (mappatura delle EST e costruzione di UniEST), la genoteca dei contig YAC e la genoteca di radiazione e 3.143 sequenze EST, ottenute da genoteche di cDNA del cervello umano, furono mappati fisicamente 308 geni. 92 di questi geni furono candidati ad essere responsabili, se mutati, di patologie che mappano nella stessa regione cromosomica (candidati posizionali, capitolo 4).

Importanza attuale delle sequenze EST.

Durante le ricerche sull'espressione dei geni vengono sintetizzati e sequenziati cDNA o parte di essi (EST) da cellule umane normali o patologiche appartenenti ad individui diversi. Queste ricerche vengono fatte con finalità scientifiche diverse, comunque quando viene individuata una nuova sequenza EST, essa viene depositata nell'archivio elettronico delle sequenze espresse umane.
La conservazione delle sequenze EST ha più scopi:
a) miglioramento della conoscenza del polimorfismo umano.
Sebbene sia già nota la sequenza totale del genoma umano, essa è stata determinata utilizzando il DNA di circa 100 individui, pertanto tale sequenza non può dare informazioni sul polimorfismo dei geni della popolazione umana.
Inoltre le EST provenienti da tessuti patologici permettono di individuare le mutazioni e l'eterogeneità dei geni responsabili delle patologie.

b) identificazione dei geni.
La migliore prova dell'esistenza di un gene è la dimostrazione che esprime mRNA che codifica una proteina. Se un cDNA EST ibrida come molecola con un frammento di DNA o la sua sequenza ibrida elettronicamente con una sequenza di DNA, essa dimostra che quel DNA e quella sequenza appartengono ad un gene, anche se la proteina codificata dalle EST è ancora ignota. L'ibridazione molecolare è stata usata in passato per identificare i geni clonati mediante la strategia del gene candidato posizionale (capitolo 4). Si è cominciato ad utilizzare l'ibridazione elettronica subito dopo che è stato costituito il primo archivio elettronico delle EST, ed in particolare durante la fase finale del progetto genoma umano, per individuare nelle sequenze del DNA genomico le sequenze che appartenevano ai geni e per stabilire la loro struttura in esoni ed introni.
La ricerca ed analisi delle sequenze EST umane continuerà fino a quando saranno individuati tutti gli alleli normali e patologici di tutti i geni presenti nella popolazione umana.

In genere, l'analisi dell'espressione dei geni è fatta valutando la concentrazione degli mRNA mediante microarray od altre tecnologie automatizzate data la loro grande potenzialità e semplicità di esecuzione; tuttavia non tutti gli mRNA presenti nelle cellule sono tradotti in tutti gli stati fisiologici (appendice B), pertanto la valutazione dell'espressione di un gene deve essere confermata da una analisi della sintesi e/o della concentrazione della proteina codificata (ad esempio aminoacidi radioattivi e/o utilizzando anticorpi specifici contro la proteina). Fanno eccezione, a quanto detto sopra, i geni degli rRNA e del tRNA che esprimono solo acidi ribonucleici.

Mappa della sequenza del DNA genomico umano.
La mappa della sequenza del DNA è costituita dalla sequenza nucleotidica completa delle molecole di DNA dei 22 autosomi e dei 2 cromosomi sessuali umani (circa 3300 miliardi di basi).
La mappa della sequenza del DNA è la più accurata delle mappe fisiche perché ha una risoluzione di una base (distanza minima misurabile). Le basi sono contate iniziando dalla prima base al termine del braccio piccolo (p) del cromosoma. Questa mappa è stata completata nell’anno 2003, essa è il risultato del progetto genoma umano iniziato nel 1990.
Per ottenere la sequenza totale del genoma è stato analizzato inizialmente il DNA di pochi individui di razze diverse. Il progetto è stato concluso con due anni di anticipo grazie al miglioramento delle tecnologie del DNA ricombinante e di quelle per l'analisi della sequenza nucleotidica, ed all'impegno di molti ricercatori dedicati alla costruzione delle altre mappe del genoma umano.
La conoscenza della sequenza completa del genoma umano ha permesso di stabilire esattamente il locus fisico di tutti i geni umani (circa 30.000), dei quali solo 10.000 erano geni noti.
Il locus fisico e la sequenza dei circa 20.000 geni ignoti sono stati individuati elettronicamente mediante ibridazione elettronica con sequenze di cDNA ed EST umane ed anche di altri mammiferi.

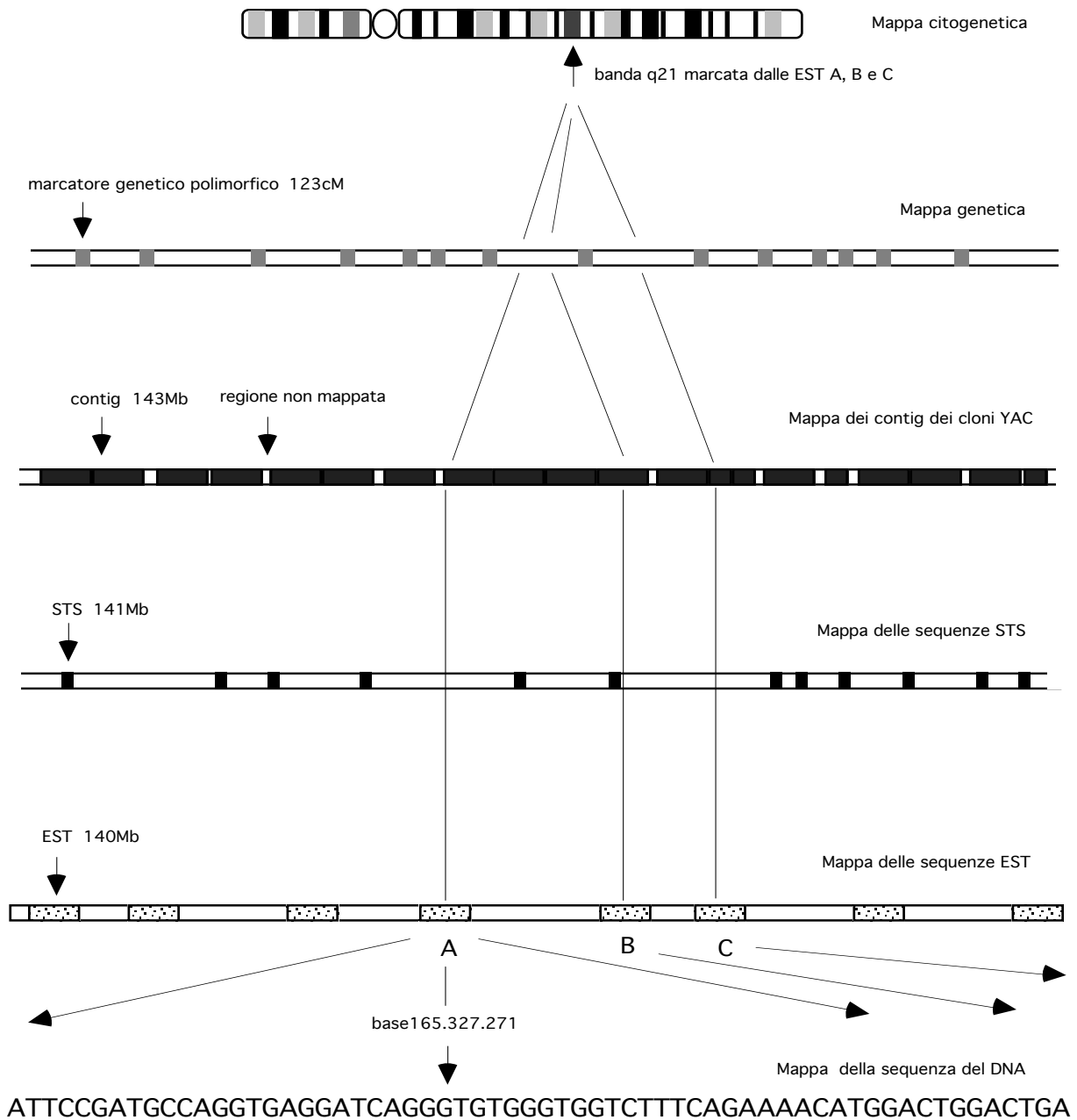



Figura 3-14. Mappe del DNA genomico umano.
A, B, C sono tre EST non polimorfiche di 3 geni diversi, marcatrici di loci nelle mappe fisiche e della banda citogenetica q21. I rapporti delle distanze fisiche tra le tre EST sono costanti nelle mappe dei contig, STS, EST e della sequenza del DNA, mentre variano nella mappa genetica di associazione che ha le distanze espresse in cM. Nel disegno la mappa della sequenza del DNA è in una scala 60 volte maggiore rispetto a quella delle EST, STS e contig, e per necessità di spazio è indicata solo una piccola porzione della sua sequenza. Nella mappa citogenetica i tre geni si trovano nella stessa banda data la minore dimensione di questa mappa, circa 1/10.000 di quella della sequenza del DNA. Le dimensioni dei marcatori non sono in scala. I marcatori al 5' del DNA cromosomico hanno indicata la loro distanza dai rispettivi marcatori più vicini al telomero del braccio p del cromosoma (la didascalia continua nella pagina successiva).

La mappa citogenetica è data dalla posizione dei geni e marcatori genetici sulle bande dei cromosomi metafasici. Le bande sono numerate sull'ideogramma di ogni cromosoma iniziando a contare dal centromero verso i due telomeri.
Il numero della banda è preceduto da "p" o "q" se la banda è rispettivamente sul braccio piccolo o su quello grande del cromosoma.

La mappa genetica è la mappa che si ottiene analizzando la frequenza di ricombinazione tra sequenze uniche polimorfiche del genoma umano: geni e marcatori genetici (VNTR, RFLP, SNP). Nella mappa, le distanze genetiche sono date in cM iniziando a contare dal marcatore più vicino al termine del braccio piccolo "p" di ogni cromosoma. 1cM equivale ad una frequenza di ricombinazione del 1% e a una distanza fisica media di 1Mb.

La mappa dei contig dei cloni YAC è la mappa di lunghi tratti di sequenza del DNA dei cromosomi umani (in figura regioni nere) detti contig, sequenze uniche del genoma umano. Il DNA delle regioni cromosomiche mappate è il 75% del totale del DNA nucleare umano. Nella mappa, le distanze fisiche sono date in numero di basi iniziando a contare dal contig più vicino al termine del braccio piccolo "p" di ogni cromosoma.

La mappa delle sequenze STS (Sequence Tagged Sites) è la mappa di sequenze uniche (polimorfiche e non polimorfiche) del genoma umano analizzabili mediante PCR. Le regioni mappate corrispondono al 95% del DNA nucleare umano. Nella mappa, le distanze fisiche sono date in numero di basi iniziando a contare dal marcatore STS più vicino al termine del braccio piccolo "p" di ogni cromosoma.

La mappa delle sequenze EST (Expressed Sequence Tags) è la mappa delle sequenze uniche (polimorfiche e non polimorfiche) del genoma che sono espresse. Nella mappa, le distanze fisiche sono date in numero di basi iniziando a contare dalla EST più vicina al termine del braccio piccolo "p" di ogni cromosoma.

La mappa della sequenza del DNA è la mappa fisica di ogni singola coppia di basi che costituiscono la sequenza del DNA di tutti i cromosomi umani. Essa ingloba tutte le sequenze marcatrici delle mappe sopra indicate. Su essa è possibile individuare la sequenza, la dimensione fisica (numero di basi) e la distanza fisica di ogni gene e di ogni tipo di marcatore genetico. Nella mappa, le distanze fisiche sono date in numero di basi iniziando a contare dalla base posta al termine del braccio piccolo "p" di ogni cromosoma. La mappa della sequenza del DNA è il completamento del progetto genoma umano.

Sono stati usati anche programmi di sequenze tipiche dei geni come le sequenze promotrici, le sequenze segnale di inizio della trascrizione e della poliadenilazione e le sequenze nelle regioni di confine tra esoni ed introni. Dalle informazioni ottenute elettronicamente, facendo alcune approssimazioni, è stato possibile calcolare in circa 30.000 il numero totale dei geni umani. Il numero non è ancora preciso dato che molti geni sono ripetuti ed alcuni di essi sono pseudogeni.
Individuata elettronicamente, la sequenza di un gene ancora ignoto nella sequenza del DNA genomico conservata nella banca dati può essere utilizzata per disegnare i primer per una PCR analitica al fine di vagliare genoteche genomiche o di cDNA ed ottenere rispettivamente la molecola del gene e del suo cDNA.
Il possesso di queste molecole permette di ricercare la funzione del gene utilizzando le tecnologie del DNA, di sintetizzare *in vitro* mediante transfezione del cDNA la proteina codificata per ricercare la sua funzione cellulare e di avere informazioni sull'attività molecolare (capitolo 2).
La conoscenza della sequenza di tutto il DNA genomico ha semplificato moltissimo la ricerca dell'espressione dei geni perché ha permesso di mettere in opera tecnologie come i microarray che analizzano simultaneamente decine di migliaia di geni (capitolo 2).
La mappa della sequenza del DNA è conservata in alcune banche dati ed è possibile fare la mappatura di una qualsiasi sequenza di DNA umano con un computer, posto in qualsiasi posto del mondo, collegandosi via internet alla pagina web "http://genome.ucsc.edu" ed usando il programma di gestione "BLAT" apribile dalla stessa pagina web. In una finestra viene scritta la sequenza di interesse e lanciando il programma dopo pochi secondi si ottiene la mappatura fisica della sequenza: numero del cromosoma e posizione subcromosomica. Se abbiamo sottoposto al programma la sequenza di un cDNA sul monitor appare disegnata anche la struttura in esoni ed introni del gene. Questo procedimento è una ibridazione elettronica o ibridazione "*in silice*" (la silice è il supporto solido dei circuiti elettronici). Alcuni autori di lingua inglese tendono a usare "*in silico*" per renderlo simile a "*in vitro*" ed "*in vivo*", ma essendo il termine latino è preferibile usarlo nella forma corretta: "*in silice*".
Da 1994, anno in cui si è iniziato a conservare le sequenze del DNA nelle banche dati e ad analizzarle mediante programmi di gestione, i modi per fare gli esperimenti di biologia sono passati da due a tre: *in vivo, in vitro ed in silice.*
Nell'ibridazione *in silice* la sequenza di interesse viene confrontata con la sequenza nucleotidica del DNAds di tutti i cromosomi ed allineata alla sequenza nucleotidica ad essa identica. Avvenuto l'allineamento esso viene disegnato sul monitor con le indicazioni della mappatura fisica: cromosoma e posizione subcromosomica fisica, locus citogenetico, inoltre la struttura del gene, delle sue varianti, le posizioni SNP, le posizioni delle sequenze EST ed STS, l'omologia con sequenze di cromosomi di topo ed altre specie. Collegandosi alla pagina web: "http: www.ncbi.nlm.nih.gov" si possono attivare in ordine le icone:
Genomes-->Eucaryotae-Genomes-->Homo-sapiens-->Chromosome-Number.

Tabella 3-3

Tipi di mappe del genoma nucleare umano		
Tipi di mappa	**Note**	**Risoluzione (distanze minime tra marcatori)**
Genetiche	**marcatori microsatelliti RFLP SNP**	**distanza media 30kb 10-300kb 1kb**
Fisiche		
Citogenetica	**bandeggio dei cromosomi**	**una banda = alcuni Mb**
Restrizione	**di enzimi con scarsi siti es. NotI**	**alcune centinaia di kb**
Radiazione	**da radiazioni di cellule ibride**	**estremi dei frammenti di cromosoma = molti Mb**
Contig dei cloni YAC	**mappato il 75% del genoma**	**2-300kb (è la dimensione degli inserti YAC)**
STS	**mappato il 95% del genoma**	**circa 100kb**
EST	**mappatura dei cDNA sulle altre mappe fisiche**	**circa 90kb**
Mappa della sequenza del DNA	**Sequenza nucleotidica del DNA dei cromosomi umani**	**1 base**

Da Strachan T. and Read A.P. (2004) Human Molecular Genetics. 3rd ed., Bios, UK, ridisegnato e modificato.

Attivando l'icona del cromosoma scelto (es. Chr 14) si hanno sul monitor le mappe della sequenza nucleotidica del DNA, di ricombinazione e citogenetica di un intero cromosoma con i collegamenti delle posizioni di ogni gene sulle tre mappe. Data la relativamente piccola dimensione del monitor, parti delle mappe possono essere espanse al fine di avere i dettagli (struttura in esoni ed introni) di ogni singolo gene. Ad ogni gene, al suo cDNA ed alla sua proteina sono stati dati codici di identificazione che possono essere utilizzati per individuarli elettronicamente negli archivi delle banche dati senza doverli ricercare utilizzando le loro sequenze e i programmi di gestione per arrivare alla stessa pagina elettronica dove sono descritti i dati del gene o della proteina.
Scrivendo il codice del gene su una apposita finestra della banca dati ed attivando il relativo programma apparirà sul monitor del computer la posizione cromosomica e sub cromosomica del gene sulle tre mappe con i geni a lui vicini ed il disegno della struttura in esoni/introni del gene. In questa banca sono stati inseriti tutti i dati (sequenze e loci) ottenuti anche prima dell'inizio del progetto

genoma. I dati di queste banche elettroniche sono continuamente aggiornati dalla immissione di nuovi dati provenienti dai laboratori di tutto il mondo. Vengono inseriti i dati di geni e marcatori neoidentificati, eliminati eventuali errori di sequenza, stabiliti nuovi collegamenti tra le varie mappe. Dato che la mappa è stata ricavata dal DNA di pochi individui, è di particolare interesse l'aggiornamento delle sequenze EST perché la loro parte codificante dà indicazioni sul polimorfismo dei geni umani delle varianti normali, di quelle subdole (appendice D) e di quelle patologiche (appendice E) della popolazione umana che è costituita da miliardi di individui.
La mappa della sequenza del DNA ha illuminato gli aspetti generali del genoma umano ed è utilissima perché è di riferimento per gli archivi elettronici contenenti gli altri tipi di mappe, le sequenze consenso, le sequenze responsabili di patologie e le sequenze delle proteine codificate.
Tuttavia, tutto ciò che riguarda gli aspetti genetici, molecolari e funzionali del singolo individuo deve essere eseguito almeno una prima volta a livello molecolare. In particolare, la ricerca dei geni responsabili delle patologie multigeniche e multifattoriali che dipendono da particolari combinazioni di alleli presenti anche in individui normali (capitolo 4 e appendice E), la ricerca di mutazioni patologiche in feti o neonati e la determinazione dell'impronta del DNA per l'identificazione di paternità, di malfattori e di cadaveri, altrimenti impossibili.
La definizione della sequenza totale del DNA genomico è stata fatta anche per lo scimpanzé, il cane, la vacca, la gallina, il ratto, il topo ed altri vertebrati, la drosofila, il nematode Cernorabditis e per alcuni tipi di lievito, virus e piante. La disponibilità della sequenza completa di genomi ha fatto nascere una nuova disciplina, la genomica comparata, che permette di studiare più dettagliatamente l'evoluzione molecolare (capitolo 1).
Il nome di mappa fisica è attribuito alle mappe citogenetiche, di radiazione, RFLP, dei contig dei cloni YAC, delle sequenze STS ed EST e di restrizione (tabella 3-3). Recentemente, essendo stata definita la mappa della sequenza del DNA che include le sequenze di tutti i geni, dei marcatori genetici e fisici, alcuni autori preferiscono distinguere le mappe del genoma umano in citogenetica, genetica e fisica intendendo per fisica solo la mappa della sequenza nucleotidica del DNA genomico.

Determinazione dell'impronta e del profilo del DNA umano

La tecnologia per l’analisi dell’impronta del DNA (DNA fingerprinting, impronta digitale del DNA) è stata inventata da A. J. Jeffrey e collaboratori nel 1985. Questi ricercatori avevano scoperto un minisatellite altamente polimorfico avente "sequenza base" GGAGGTGGGCAGGAXG (X può essere una qualsiasi delle 4 basi) e che ripetizioni diverse della sequenza base costituivano più specie molecolari di VNTR, disposte con alte percentuali di eterozigosi su molti loci, presenti su tutti i cromosomi.

Gli stessi ricercatori fecero una analisi Southern di DNA genomico umano digerito con un enzima di restrizione HinfI ed utilizzarono come sonda la sequenza base del minisatellite. La sonda ibridava su tutti i tandem dei frammenti ottenuti per digestione che provenivano da circa 60 loci diversi. L'autoradiografia mostrava circa 35 bande aventi una disposizione unica per ogni individuo analizzato, cioè una impronta dei DNA specifica dell'individuo (figura 3-15a). Il minor numero di bande rispetto al numero di loci è imputato alla comigrazione nell'elettroforesi di frammenti aventi lo stesso numero di basi, ma provenienti da loci diversi.
La probabilità che le disposizioni delle bande di due individui non parenti coincidano è $3x10^{-11}$ con una sonda e di 10^{-20} con due sonde diverse (che si legano a due sequenze base diverse). Teoricamente si assume che solo in una popolazione rispettivamente di 300 miliardi e 100 miliardi di miliardi possono capitare due individui con la stessa combinazione di bande.
Tuttavia questa tecnica è stata abbandonata principalmente per tre motivi:
1. Occorrono alcuni microgrammi di DNA genomico, quantità relativamente grande che richiede di estrarre il DNA da molte cellule.
2. L’impronta del DNA ottenuta con sonde multilocus non indica il locus delle bande né l’origine paterna o materna di una data banda.
3. Una certa difficoltà ad analizzare l’autoradiografia dato l’alto numero di bande da esaminare e confrontare in base alla loro posizione ed intensità.
La tecnica per l’analisi dell’impronta del DNA (DNA fingerprinting) è stata sostituita con una nuova detta “Profilo del DNA” (DNA profiling) che analizza, con la tecnica PCR analitica, un gruppo di microsatelliti altamente polimorfici, non geneticamente associati (ricombinanti ad ogni meiosi), costituiti da ripetizioni di 2, 3 o 4 basi e disposti su loci diversi. Di ogni singolo locus sono individuati i due alleli di un dato individuo ed effettuando l'analisi degli alleli dello stesso locus sul DNA dei suoi genitori si ha l'indicazione dell'origine paterna o materna di ciascun allele (figura 3-16).
Poiché il DNA da analizzare viene amplificato durante l'analisi PCR, sono sufficienti quantità piccolissime di DNA, come quelle presenti in un capello, nella saliva lasciata in una sigaretta, in una goccia secca di sangue o il DNA di una singola cellula.
L’analisi dei microsatelliti di 10-15 loci diversi è sufficiente ad identificare inequivocabilmente un individuo permettendo di distinguere il profilo del suo DNA da quello di un qualsiasi estraneo ma anche da quelli delle sorelle e dei fratelli, dei genitori, dei nonni, degli zii, cugini e più lontani parenti. Il confronto dei vari profili permette di stabilire il grado di parentela tra un individuo ed i suoi parenti, anche lontani, pertanto essi possono essere utilizzati per costruire l’esatto albero genealogico delle famiglie.
Il profilo del DNA, che con 10-15 loci polimorfici, individua geneticamente un individuo è analogo al profilo di un volto i cui pochi tratti somatici permettono di identificare un individuo.

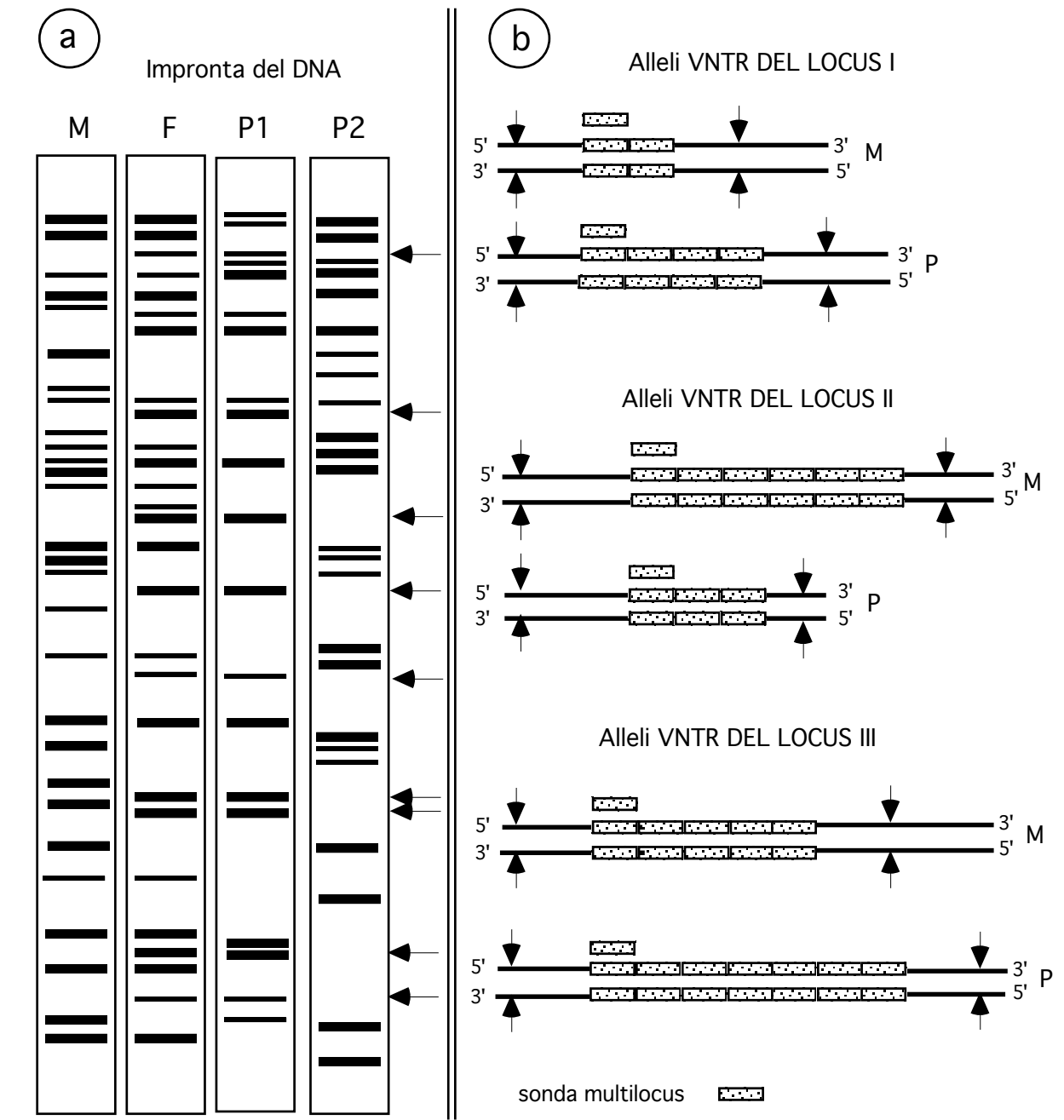



Figura 3-15. a) Impronta dei DNA mediante autoradiografia dell'elettroforesi di DNA umano digerito con l'enzima di restrizione HinfI ed analizzato con un'analisi Southern utilizzando una sonda multilocus. M=madre; F=figlio; P1 e P2 padri da identificare. Tutte le bande della traccia devono essere presenti nella traccia della madre o del padre. Se nel proposto padre è assente una banda che è presente nel figlio, essa lo esclude dalla paternità. P1 è il padre naturale. Le frecce orizzontali indicano alcune bande assenti nelle tracce M e P2 e presenti nelle tracce F e P1. Le bande più spesse includono frammenti di DNA delle stesse dimensioni provenienti da loci diversi.
b) Sono indicati gli alleli di tre distinti loci di una sequenza VNTR aventi la stessa sequenza unitaria (GGAGGTGGGCAGGAXG). Tagliando con lo stesso enzima di restrizione (i punti di taglio sono indicati dalle frecce verticali) si hanno frammenti di lunghezza diversa che possono essere rivelati con una sonda multilocus costituita dalla sequenza unitaria (o parte di essa) che non discriminando i loci (come per la sonda mostrata in a) produrrà 6 bande diverse (1, 2, 3, 4, 6 e 7 ripetizioni in tandem). La lunghezza dei frammenti è in relazione alle ripetizioni in tandem e alle due posizioni dei siti di taglio che, in uno stesso locus, sono identiche. In figura la posizione di ibridazione della sonda è solo indicativa, in realtà essa può legarsi a qualsiasi sequenza unitaria e, se in eccesso, a tutte quelle di un minisatellite.

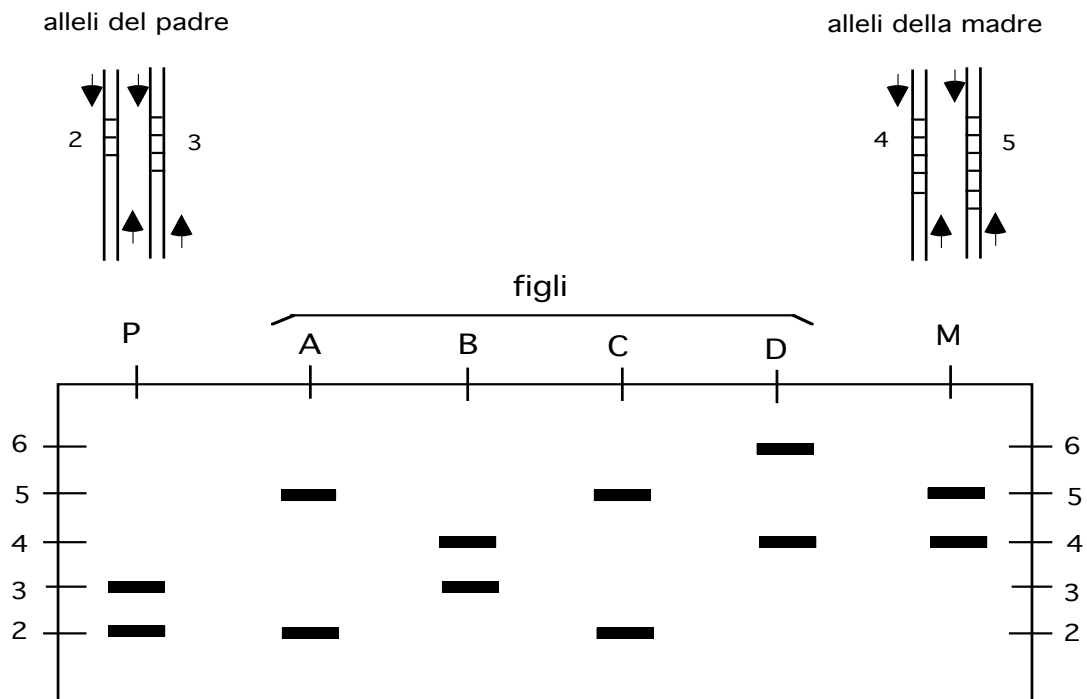



Figura 3-16. Analisi mediante la tecnica PCR del locus di un microsatellite per la determinazione del profilo del DNA. La banda contenente un microsatellite con 6 ripetizioni in tandem indica che il figlio D non ha lo stesso padre degli altri figli. Le frecce indicano i primer che si associano a regioni uniche nel genoma che marcano il locus del microsatellite.

Per escludere l'identità tra due profili di DNA è sufficiente che una singola banda di un singolo locus di un profilo abbia una migrazione diversa dalla banda ad essa omologa dell'altro profilo.
Per escludere la paternità, è sufficiente che uno dei due alleli paterni dei 10-15 loci analizzati non sia presente negli stessi loci del DNA del figlio, cioè che in un dato locus del figlio ambedue le bande nell'elettroforesi migrino diversamente da quelle del padre e della madre (figura 3-16). Con lo stesso criterio è esclusa la responsabilità in un crimine da parte di un individuo, quando il profilo del suo DNA risulta diverso da quello del DNA delle cellule del sangue trovato sulla scena del crimine.
Per provare l'identità tra due profili occorre che tutte le bande degli alleli degli stessi loci abbiano la stessa migrazione.
Per provare la paternità occorre che uno dei due alleli paterni dei 10-15 loci analizzati sia sempre presente negli stessi loci del DNA del figlio, cioè che in tutti i loci del DNA del figlio una delle due bande migri all'elettroforesi in maniera identica ad una del padre e l'altra in maniera identica a quella della madre certa.
In genere è sufficiente la coincidenza delle bande di alleli (del padre e del figlio) di 10-15 loci per avere la sicurezza che il padre sia il genitore naturale del figlio. La possibilità che, oltre all'individuo analizzato, ne possa esistere un altro con la stessa combinazione di alleli è di 1 su 100 miliardi, in una popolazione umana di 10 miliardi.
Il profilo del DNA di un individuo deve essere identico al profilo del DNA dei reperti biologici trovati nella scena del crimine per poterlo sospettare/imputare di quel crimine. Le bande di tutti i loci dei due DNA devono coincidere.
Sebbene la definizione del profilo del DNA sia ritenuta valida e quindi accettata dai tribunali di tutto il mondo ed in genere anche dal genitore sospettato o sospettoso, tuttavia non si può escludere (almeno teoricamente) che tra i viventi esistano due individui con la stessa combinazione di bande (alleli) nei 10-15 loci. In altre parole, sebbene il profilo del DNA mostri il figlio quale figlio del padre anagrafico, il figlio sarebbe stato generato da un altro padre (padre biologico) se la madre avesse avuto una relazione con un uomo che per caso aveva la stessa combinazione di alleli del proprio marito. Molto più improbabile che vincere al superenalotto con una giocata minima! Salvo che la donna abbia fatto fare il profilo del DNA dell'amante e poi sia andata avanti tranquilla, ma si assume che questa possibilità non esista.

Tuttavia è stato verificato che circa l'1% dei profili del DNA sono errati per errori di valutazione dell'allineamento delle bande fatti dai tecnici.
La tecnologia del profilo del DNA è stata ulteriormente semplificata ed automatizzata. La PCR è effettuata in una unica incubazione contenente le coppie di primer per tutti i loci dei microsatelliti di interesse ed alcune coppie di primer sono legate covalentemente a fluorocromi diversi al fine di distinguere gli amplificati di microsatelliti appartenenti a loci diversi che possono avere la stessa migrazione elettroforetica. Le varie specie molecolari di DNA amplificato

sono frazionate in base al loro peso molecolare mediante elettroforesi capillare ed identificati in base alla lunghezza d'onda della loro fluorescenza.
Quando necessario, per l’identificazione di individui sono anche utilizzati il polimorfismo del DNA dei cromosomi sessuali X ed Y che sono rispettivamente di origine materna e paterna.
L’impronta del DNA è utilizzata anche per definire geneticamente animali e piante al fine di individuare specie e razze importanti per scopi alimentari, estetici e sportivi. Con il termine “impronta del DNA” viene indicata anche la tecnica di identificazione di batteri, cioè la loro appartenenza a ceppi diversi di una stessa specie. I batteri non hanno sequenze VNTR e l'impronta del DNA viene effettuata utilizzando sequenze conservate e disperse nel genoma batterico. Quindi "impronta del DNA" ha acquistato il significato più generale di identificazione genetica di individui di una data specie sulla base di sequenze specifiche.
Il profilo del DNA permette l'identificazione veloce e sicura di un individuo sulla base di una specifica combinazione di un gruppo di sequenze non codificanti e ripetute di DNA. Tuttavia la vera individualità genetica di ogni essere vivente è definita dalla combinazione degli alleli dei geni che ha ricevuto dai genitori.

Meccanismi molecolari di formazione del polimorfismo delle sequenze VNTR

Nelle figure 3-17 e 3-18 sono mostrati due meccanismi di formazione del polimorfismo delle sequenze VNTR: slittamento dei filamenti del DNA (slittamento intracromatidico) durante la sintesi, crossing over ineguale e scambio ineguale tra cromatidi fratelli. Queste possibilità non incrinano l’affidabilità della tecnica del DNA fingerprinting perché i loci che hanno sequenze VNTR che presentano da una generazione all’altra anche una bassissima frequenza di variazione nel tandem (uno o pochi individui nel mondo) non vengono utilizzati per l’analisi. Quando si presentano nuovi casi vengono prontamente comunicati alla comunità scientifica mondiale.
I meccanismi di generazione di polimorfismo di microsatelliti con sequenza unitaria di tre basi sono responsabili di alcune patologie (vedere appendice E).

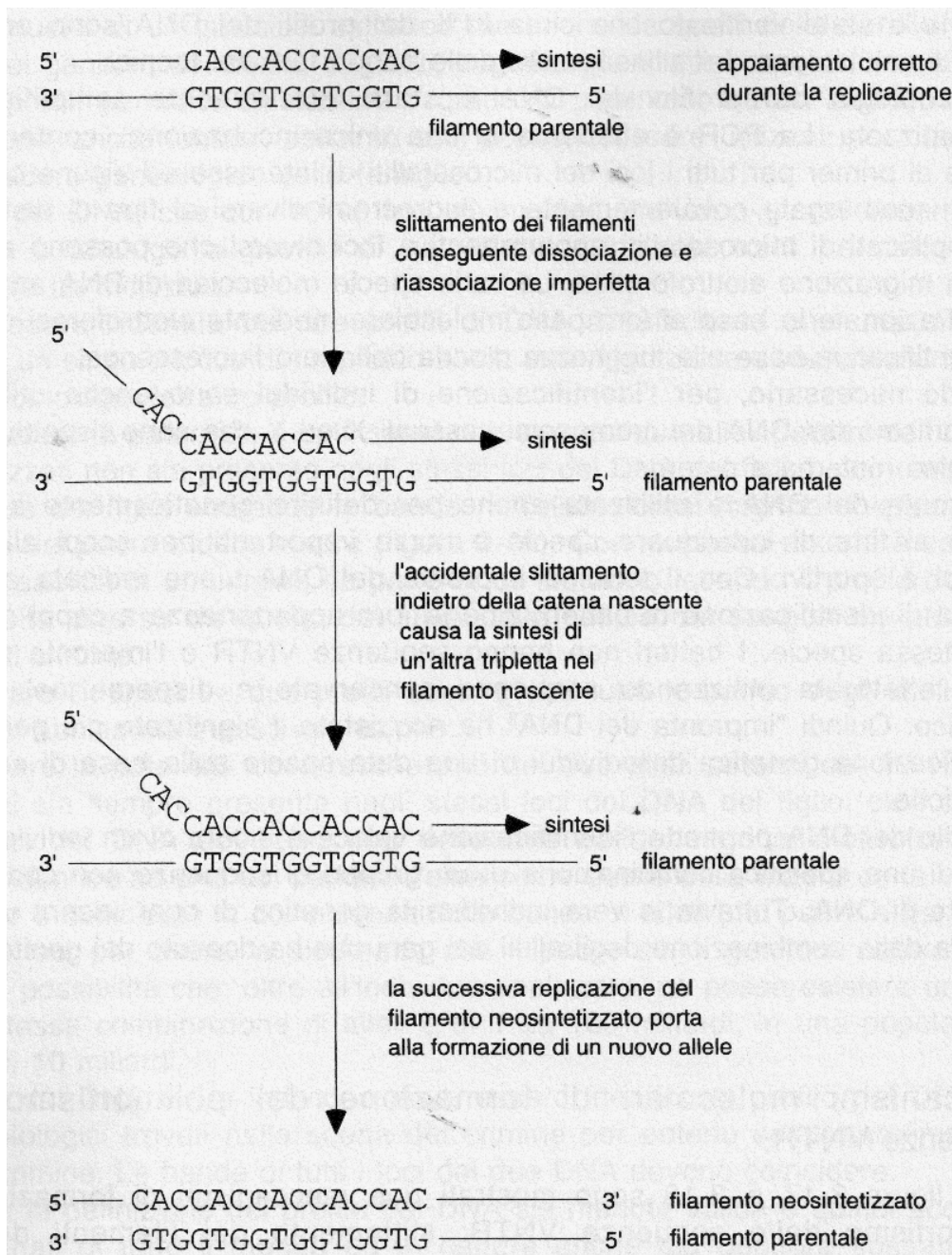



Figura 3-17. Meccanismo di formazione del polimorfismo VNTR per slittamento dei filamenti, nascente e parentale-stampo, durante la replicazione del DNA. L'altro filamento parentale (non mostrato in figura) è ricopiato come era (o subisce anch'esso uno slittamento). Terminate la replicazione del DNA e la divisione cellulare, una cellula figlia avrà un tandem con cinque triplette e l'altra cellula figlia avrà nello stesso locus del cromosoma omologo un tandem con quattro triplette (se non ha subito anch'essa uno slittamento durante la sintesi del DNA). I due filamenti possono avere anche uno slittamento in senso inverso e le triplette da quattro divenire tre (vedere figura E-8). Si assume che questo meccanismo sia responsabile della formazione dell'alto polimorfismo dei microsatelliti (ridisegnato e modificato da Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK).

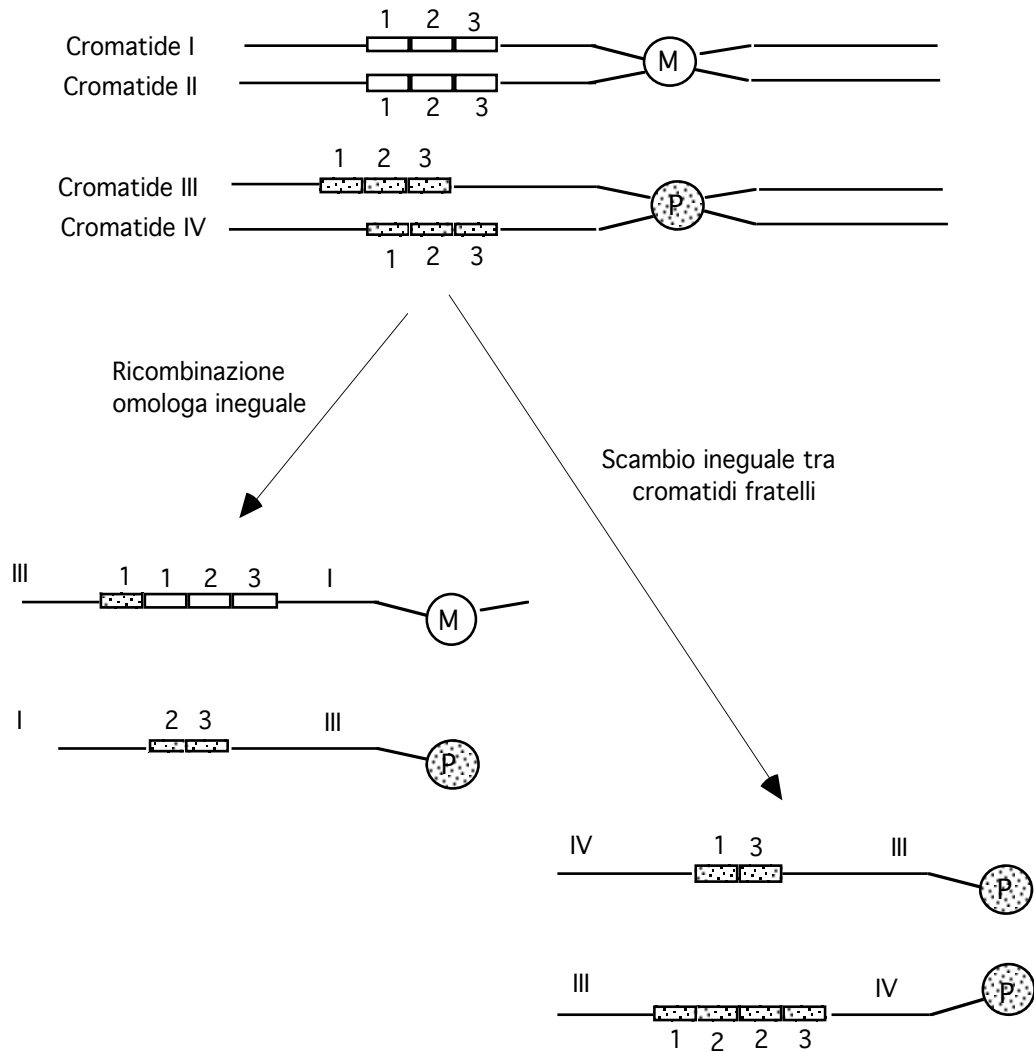



Figura 3-18. Formazione del polimorfismo delle sequenze VNTR per crossing over ineguale o per scambio ineguale di cromatidi fratelli. Si assume che questo meccanismo sia responsabile del polimorfismo dei minisatelliti. La frequenza di formazione di nuovi alleli è inversamente proporzionale alla lunghezza della sequenza unitaria. M = Materno, P = Paterno.
(ridisegnato e modificato da Strachan T. and Read A.P. (1996) Human Molecular Genetics. Bios, UK).

Guarda le cose che vedono tutti e pensa differentemente.
Albert Szent-Gyorgyi, Premio Nobel per la Medicina.

Capitolo 4.

Strategie per la clonazione dei geni e progetto genoma umano

Dopo la messa in opera delle tecnologie del DNA, la strategia della clonazione funzionale dei geni è stata la strategia più utilizzata perché erano note le caratteristiche molecolari di molte proteine normali e di alcune di esse era nota la forma mutata responsabile di una patologia monogenica (capitolo 1). In questo modo furono clonati i geni della suscettibilità a circa 400 patologie monogeniche. Tuttavia l'uso di questa strategia è diminuito nel tempo per esaurimento delle proteine note e perché la ricerca e la purificazione di nuove proteine normali e patologiche è un processo molto lento rispetto alla clonazione del gene, una volta che sia nota la struttura primaria o l'attività molecolare della proteina da esso codificata. Dal 1986, con la messa in opera della strategia della clonazione posizionale dei geni sono stati identificati geni responsabili di malattie genetiche delle quali era nota solo la posizione subcromosomica (locus) stabilita utilizzando marcatori genetici associati alla malattia, e nulla si conosceva della proteina da essi codificata. Alla fine del 1993 con la strategia posizionale, erano stati clonati 19 geni della suscettibilità a patologie monogeniche. La strategia della clonazione posizionale dei geni è considerata da alcuni autori un altro aspetto della genetica inversa perché si opera direttamente sul gene senza conoscerne la proteina codificata (dal gene alla proteina e non dalla proteina al gene), anche se l'informazione della presenza del gene proviene dal fenotipo alterato. L'applicazione di questa strategia è diminuita nel tempo per esaurimento delle patologie delle quali era noto il locus.

In assenza di dati precisi sulla struttura primaria, sulla attività molecolare o sulla funzione cellulare di una proteina responsabile di una patologia monogenica o della mappatura della patologia monogenica di interesse, sono state utilizzate nuove strategie dette del "gene candidato" ad essere responsabile della patologia.

In alcuni casi è stato possibile clonare e mappare un gene utilizzando dati di un gene omologo (ortogene) di un'altra specie. L'enzima glicerolo-cinasi umana è stato clonato utilizzando una sequenza EST umana, della quale era ignoto il gene che l’aveva espressa, ma che aveva una sequenza simile a quella dell'enzima glicerolo-cinasi batterica.

Al fine di avere la visione completa dei geni umani nel 1990 fu varato il progetto genoma umano: la determinazione della sequenza totale delle 24 molecole DNA dei cromosomi umani. La sequenza è stata completata nel 2003. Questa è stata la più grande opera di biologia molecolare di tutti i tempi.

Strategie per la clonazione dei geni

Clonazione funzionale dei geni

Conoscendo parte della sequenza o un metodo specifico di analisi di una proteina è possibile clonare il relativo gene.

```
Attività molecolare ---> proteina --->   Sequenza  ----->    sintesi di oligonucleotidi
                                       aminoacidica                 degenerati


saggio funzionale o     anticorpo
di complementazione



    vaglio della -------> sequenza --------> vaglio della -------> mappatura fisica
  genoteca di cDNA        del cDNA       genoteca genomica           del gene
```

Clonazione dei geni candidati funzionali

Quando è possibile ipotizzare che la mutazione di una proteina possa provocare le stesse alterazioni di una patologia monogenica, è possibile identificare e clonare il gene responsabile della suscettibilità alla patologia.

```
Clonazione ----> mappatura ----> dimostrazione ----> dimostrazione ---> l'allele mutato
funzionale       genetica        che un allele        che l'allele        codifica una
del gene         del gene        è associato           associato è          proteina
                                 alla patologia          mutato             inattiva
```

Clonazione posizionale dei geni

Conoscendo i marcatori genetici di una patologia monogenica che mappa in una regione cromosomica includente un solo gene, è possibile clonare il gene della suscettibilità a quella patologia.

```
Vaglio di una --------> camminare ------> sequenza -------> dimostrazione che un allele
genoteca genomica     sul cromosoma       del gene            del gene è mutato ed
mediante la                               candidato              associato ad una
tecnologia PCR                                              patologia in più famiglie
specifica per                                                 di aree geografiche e/o
il marcatore                                                       etnie diverse
```

Clonazione dei geni candidati posizionali

Conoscendo i marcatori genetici di una patologia monogenica che mappa in una regione cromosomica includente più geni candidati, è possibile clonare il gene responsabile della suscettibilità a quella patologia.

```
Vaglio di una ---------> camminare ------> sequenza -----> dimostrazione che l'allele
genoteca genomica      sul cromosoma       dei geni          di uno solo dei geni è
mediante PCR                               candidati        mutato ed associato alla
specifica per                                               patologia in più famiglie
il marcatore                                                 di aree geografiche e/o
                                                                  etnie diverse
```

Clonazione dei geni dei fenotipi complessi normali e patologici

Analizzando un alto numero di marcatori genetici presenti nel DNA di tutti i cromosomi è possibile individuare i loci dei geni candidati funzionali a determinare un fenotipo complesso o a causare una patologia complessa.

```
Determinazione del ------------> Determinazione in silice -------> Identificazione in
profilo degli alleli          della sequenza delle regioni          silice dei geni
dei marcatori genetici di        cromosomiche associate           candidati funzionali
tutto il genoma mediante       a un fenotipo complesso o         presenti nelle regioni
scansione posizionale          ad una patologia complessa         cromosomiche mappate
```

Strategia della clonazione funzionale dei geni

La strategia della clonazione funzionale permette di clonare un gene conoscendo una caratteristica strutturale o un metodo specifico di analisi della proteina da esso codificata.

Cenni di biochimica tradizionale

La ricerca biochimica, che ha avuto un grande impulso a partire dagli anni '40, ha individuato e purificato molte proteine realizzando tecnologie per saggiare la loro attività molecolare, la funzione cellulare e per purificarle. In genere la ricerca iniziava con lo studio di manifestazioni fisiologiche cellulari come la digestione degli alimenti, la degradazione del glucosio, la sintesi e degradazione del glicogeno, la contrazione delle fibre muscolari o dallo studio di alterazioni patologiche delle stesse funzioni cellulari. Queste conoscenze, favorendo la clonazione dei geni, hanno molto contribuito allo sviluppo della biologia molecolare.

Alcuni autori infatti considerano gli anni '40, il periodo di inizio della biologia molecolare perché in quegli anni la Fondazione Rockefeller, che finanziava la ricerca internazionale, stabilì che avrebbe finanziato solamente i progetti di ricerca biologica che utilizzavano tecnologie chimiche e fisiche.

Di seguito viene descritto brevemente come il biochimico analizzando una funzione cellulare alterata riusciva ad individuare la proteina responsabile dell'alterazione. Lo scopo è di mostrare come partendo da una alterazione funzionale ed utilizzando tecniche biochimiche si arrivi ad individuare la proteina alterata e poi con le tecniche del DNA al gene che la codifica.

In alcuni stati patologici viene incrementata la concentrazione di uno o più metaboliti intermedi che sono poco concentrati nelle cellule normali e questi composti possono riversarsi nel sangue e, se in eccesso, dal sangue nelle urine. Quando si osservava che un paziente aveva nelle urine (o nel sangue o in altri distretti) un composto che era scarsamente presente o non presente nelle urine degli individui normali (esempio fenilchetonuria, appendice E) si cercava di identificarlo per poi risalire all'enzima che poteva aver catalizzato la reazione della sua sintesi e a quello che poteva averlo come substrato.

L'incremento di concentrazione di metaboliti, causato dalle patologie, permetteva di osservare la loro presenza con le tecnologie di quel tempo, che risultavano non sufficientemente sensibili per rivelare gli stessi metaboliti nelle concentrazioni normali perché troppo basse. Le carenze di metaboliti in genere sono più difficili da individuare, specialmente se lo stesso metabolita è scarsamente concentrato anche nelle cellule normali. Tuttavia, quando si riesce ad individuare una differenza tra normale e patologico, come la variazione di concentrazione di un metabolita, essa viene sempre indagata al fine di meglio caratterizzare la patologia per arrivare ad individuare le sue cause (eziologia), il meccanismo molecolare della sua formazione (patogenesi) e le alterazioni molecolari responsabili dei sintomi clinici percepibili direttamente dai nostri

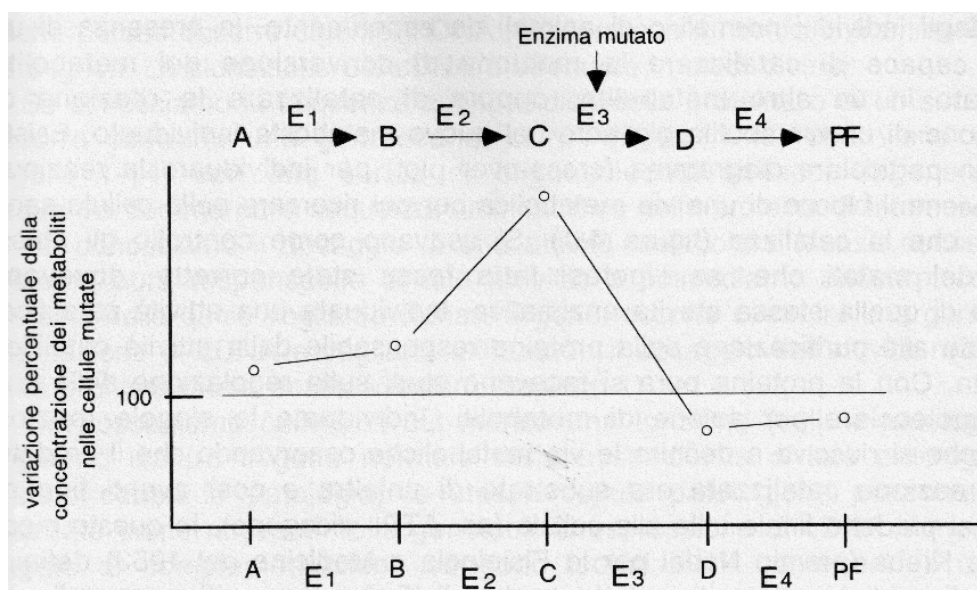



Figura 4-1. Cross-over plot (diagramma di passaggio sopra). Sull'ascissa vengono riportati i valori delle variazioni percentuali delle concentrazioni dei metaboliti nelle cellule mutate nell'enzima E3. I valori delle concentrazioni dei metaboliti nelle cellule normali sono stati normalizzati al valore di 100. Il passaggio da valore positivo a negativo di variazione percentuale indica che il blocco metabolico è tra il metabolita C e quello D e ciò può dare indicazioni sul tipo di reazione catalizzata dall'E3 e quindi sul dosaggio della relativa attività catalitica (ridisegnato e modificato da F. Gabrielli, Gli Enzimi. Piccin, Padova, 1977).

---

organi di senso o per mezzo di tecnologie che non sono altro che mezzi per rendere percepibili ai nostri organi di senso fenomeni altrimenti non percepibili.
Una patologia può essere causata da una proteina geneticamente alterata o da una proteina fenotipicamente alterata da radiazioni, da agenti infettivi, da molecole provenienti dall'ambiente o assunte con gli alimenti (appendice E).
Le patologie sono causate da alterazioni di proteine con l'eccezione delle alterazioni causate dall'ossidazione dei fosfolipidi di membrana.
Le patologie genetiche possono risultare da sintesi eccessive o insufficienti di una proteina normale o da sintesi di una proteina alterata nella sequenza (mutata). Queste alterazioni delle proteine causano patologie provocando l'alterazione della concentrazione, nelle cellule e nei liquidi biologici, di metaboliti normali senza produrre alterazioni della molecola degli stessi metaboliti (appendice E).
Le analisi biochimiche portano a definire la formula di struttura del metabolita, a riconoscerlo come possibile intermedio metabolico ed a stabilire la sua possibile conversione, mediante una singola reazione covalente, in un metabolita già noto di una via metabolica. Basandoci sulla conoscenza che nella cellula tutte le reazioni covalenti sono catalizzate da enzimi, si cercava nei tessuti degli individui normali o di animali da esperimento, la presenza di un enzima capace di catalizzare la reazione di conversione del metabolita individuato in un altro metabolita, oppure di catalizzare la reazione di conversione di un metabolita già noto nel nuovo metabolita individuato. Esiste

anche un particolare diagramma (cross-over plot) per individuare la reazione dove avviene il blocco di una via metabolica per poi ricercare nelle cellule sane l'enzima che la catalizza (figura 4-1). Si usavano come controllo gli stessi tessuti dei malati che, se l'ipotesi fatta fosse stata corretta, dovevano mancare di quella stessa attività enzimatica. Individuata una attività catalitica, si procedeva alla purificazione della proteina responsabile della attività catalitica osservata. Con la proteina pura si facevano studi sulla regolazione della sua attività molecolare per azione di metaboliti. Individuate le singole reazioni enzimatiche si riusciva a definire le vie metaboliche osservando che il prodotto di una reazione catalizzata era substrato di un'altra e così avanti fino ad arrivare al prodotto finale utile alla cellula (es. ATP, glicogeno). In questo modo Sir Hans Krebs (premio Nobel per la Fisiologia e Medicina nel 1953) definì il ciclo degli acidi tricarbossilici, detto anche di Krebs. La purificazione di una proteina avveniva dosando l'attività molecolare della stessa, inizialmente negli omogenati grezzi (contenenti tutte le proteine cellulari), poi in soluzioni sempre più ricche della proteina ricercata, perché purificata dagli omogenati di cellule normali mediante vari tipi di tecnologie biochimiche (precipitazioni frazionate con sali, cromatografie, elettroforesi, ecc.) utilizzate in successione per ottenere frazioni arricchite di proteina di interesse (individuata mediante il dosaggio della sua attività molecolare) e frazioni da eliminare contenenti le altre proteine (contaminanti) che non avevano l'attività molecolare di interesse. Avendo la proteina pura si poteva studiarne più dettagliatamente l'attività molecolare, la sua funzione cellulare, si poteva determinare la sequenza aminoacidica ed analizzare la sua struttura tridimensionale.

Strategia della clonazione funzionale

Negli anni '70 i biologi molecolari avevano a disposizione i dati sull'attività molecolare e sulla funzione cellulare di molte proteine pure e di un relativamente piccolo numero di esse anche la sequenza aminoacidica. Tutto ciò fornito dal precedente e paziente lavoro dei biochimici.
La strategia della clonazione funzionale permette di clonare un gene utilizzando le caratteristiche della molecola o dell'attività molecolare della proteina da esso codificata. Si inizia clonando il cDNA del gene di interesse, e non direttamente il gene stesso, perché le genoteche di cDNA, avendo meno cloni, sono più semplici da costruire e da vagliare di quelle genomiche e perché si possono costruire anche genoteche di espressione di cDNA (utili per clonare geni quando non sia nota la sequenza aminoacidica della proteina di interesse), mentre non si costruiscono genoteche genomiche di espressione, perché l'espressione *in vitro* di tutti i geni del genoma è praticamente impossibile, ed anche perché occorre digerire il DNA nucleare con il rischio di frammentare dei geni ignoti. La clonazione del cDNA è effettuata in modi diversi:
1. Se conosciamo la sequenza aminoacidica della proteina o di un suo peptide occorre costruire una genoteca di cDNA (figura 1-13) che viene vagliata (figura 1-14) con una sonda di oligonucleotidi sintetici degenerati con sequenza dedotta dalla sequenza aminoacidica della proteina di interesse.

2. Se conosciamo il dosaggio di una attività molecolare (senza conoscere la proteina pura responsabile di tale attività) o possediamo anticorpi contro la proteina della quale vogliamo clonare il gene, occorre costruire una genoteca di espressione di cDNA e vagliarla utilizzando il dosaggio dell'attività molecolare della proteina o una analisi Western (capitolo 1).
3. Se possediamo cellule in cui manca una funzione (cellule patologiche) e vogliamo isolare il gene normale della suscettibilità a tale patologia (che mutato causa la patologia) le stesse cellule patologiche possono essere utilizzate come riceventi della genoteca di espressione di cDNA e mediante il saggio di complementazione vengono isolati cloni contenenti il costrutto (vettore-cDNA) che complementa le cellule restaurando la funzione normale (capitolo 1). Clonato il cDNA si determina la sua sequenza nucleotidica (figura 1-10), nella sequenza viene individuato il quadro di lettura aperto (orf) e da esso viene dedotta la sequenza aminoacidica dell'intera proteina (figura 1-15).
4. Il gene viene clonato vagliando una genoteca genomica utilizzando una PCR analitica specifica per il cDNA di interesse oppure utilizzando come sonda il cDNA o parte di esso. Poi il gene è mappato fisicamente utilizzando la mappa fisica delle sequenze STS. Al fine di mappare geneticamente il gene, sono ricercati i suoi marcatori genetici microsatelliti all'interno degli introni e camminando sul cromosoma in regioni vicine al 5' e 3' del gene.
Definire il locus (fisico e genetico) di un gene è importante perché esso può coincidere con quello di una malattia di cui si conosce il locus ma non il gene, né l’attività biologica della proteina da esso codificata.
In genere l'identificazione del gene viene confermata da altri studi, ad esempio il cDNA è transfettato in cellule batteriche o di lievito e la proteina prodotta viene sottoposta alle analisi molecolari (dosaggio dell'attività molecolare, sensibilità ad effettori, ecc.) che sono state usate per caratterizzare la proteina nativa pura. Vengono prodotti anticorpi immunizzando animali con la proteina sintetica al fine di verificare la presenza della proteina nei vari tessuti umani. Il confronto della sequenza del cDNA con quella del gene permette di identificare la struttura in esoni ed introni del gene e di verificare se la proteina è prodotta da uno splicing alternativo quando la sua sequenza è codificata da un cDNA che non include tutti gli esoni.
Determinate le sequenze del gene e della sua proteina, esse vengono confrontate con quelle conservate nelle banche dati (vedere figure 1-17 e 1-18). L’indagine elettronica permette di verificare se gene e proteina di interesse siano già stati clonati oppure se la proteina di interesse appartiene ad una famiglia di proteine. La similarità strutturale con altre proteine può fornire informazioni anche sulla attività molecolare della proteina di interesse, quando si riscontri la presenza di domini con attività molecolari specifiche (figura 1-19).

Strategia della clonazione dei geni candidati funzionali

Quando si conosce l'attività molecolare e la funzione cellulare di una proteina e si può ipotizzare che una sua alterazione genetica possa provocare gli stessi sintomi (manifestazioni di alterazioni morfologiche o funzionali) di una patologia monogenica di interesse, il gene che la codifica è definito candidato ad essere il gene della suscettibilità a quella patologia monogenica, cioè se mutato è responsabile della patologia.
I dati della proteina, utili a candidare il gene, possono essere: la proteina partecipa ad una via metabolica che appare alterata nella patologia (carenza/eccesso di un metabolita)(figura 4-1); la proteina svolge un ruolo nell'organogenesi e la patologia è un'alterazione della formazione di un organo; la proteina ha la stessa localizzazione tissulare (es. fegato) e la stessa localizzazione subcellulare della patologia monogenica (es. la proteina è associata alla membrana plasmatica e la patologia si manifesta con alterazioni della membrana plasmatica). Assumendo che la proteina alterata possa essere responsabile della patologia monogenica, il gene normale che la codifica diviene il gene candidato della suscettibilità alla patologia e le caratteristiche di attività molecolare e funzione cellulare della proteina possono essere utilizzate per isolare il gene mediante clonazione funzionale.
Per stabilire con certezza che il gene clonato come candidato funzionale sia quello della suscettibilità alla patologia studiata occorre eseguire delle analisi genetico-molecolari per verificare che un allele del gene candidato è associato alla patologia e che quell'allele è mutato e codifica una proteina con scarsa o nulla attività molecolare (vedere dopo gli schemi delle strategie di clonazione e la patologia MODY). Conoscendo la sequenza o la funzione della proteina candidata si opera la clonazione funzionale del gene che la codifica, si ricercano i suoi marcatori genetici microsatelliti all'interno degli introni e camminando sul cromosoma in regioni vicine al 5' e 3' del gene. Quindi, utilizzando un marcatore del gene, si verifica se in una famiglia con alta incidenza di una patologia monogenica, un particolare allele del gene candidato sia presente nei membri malati ed assente in quelli sani. In caso positivo, il gene candidato funzionale diviene anche candidato posizionale, e si ha l'indicazione che in quella famiglia non avviene la segregazione tra patologia ed un dato allele del gene di interesse. Tuttavia il gene di interesse potrebbe essere non patologico e l'allele associato alla patologia una variante normale (in questo caso il gene patologico sarebbe un altro gene facente parte dello stesso aplotipo). Occorre verificare che l'allele esclusivamente associato ai membri malati codifichi una proteina inattiva rispetto alla proteina prodotta dall'allele presente nei membri sani. In caso affermativo si ha la prova che l'allele associato alla patologia produce una proteina inattiva e quindi può essere responsabile della patologia. La verifica che l'allele produce una proteina inattiva può essere la semplice constatazione di segnali di stop alla traduzione posti all'inizio del mRNA, comunque si utilizzano i cDNA di tutti gli alleli varianti per sintetizzare le

relative proteine e si verifica con un dosaggio specifico se le proteine hanno o non hanno attività molecolare. Un'ulteriore ed importante conferma della individuazione del gene della suscettibilità alla patologia viene dalle analisi genetico-molecolari che individuano l'allele patologico in membri portatori della patologia appartenenti ad altre famiglie di aree geografiche o etnie diverse.
Un ulteriore studio della patologia può essere fatto ricercando in un'altra specie il gene omologo a quello di interesse, manipolarlo con le tecnologie descritte nei capitolo 3 per verificare se nell'animale il gene mutato produca effetti simili a quelli osservati nell'uomo e quindi studiarlo tramite esperimenti non possibili sull'uomo. Purtroppo (non per i poveri topolini) talvolta gli effetti negli animali sono diversi da quelli osservati nell'uomo.

Strategia della clonazione posizionale dei geni

La clonazione posizionale (o di posizione) è la strategia usata per clonare un gene partendo dalla sola conoscenza della localizzazione subcromosomica (locus) del suo fenotipo patologico. Quando la posizione subcromosomica del gene di interesse è associata ad una regione di DNA cromosomico marcata da più marcatori genetici che definiscono una regione cromosomica in cui è incluso un solo gene, la clonazione è definita clonazione posizionale (se include più geni, la clonazione è detta clonazione dei geni candidati posizionali, vedere dopo).
Una regola dettata dalla pratica suggerisce che, se analizzando almeno 100 pazienti generati da gameti provenienti da meiosi informative si osserva una sola ricombinazione, la distanza tra il marcatore e la patologia é da considerarsi breve al fine di intraprendere la clonazione del gene (1 ricombinazione su cento meiosi informative corrisponde a un 1% di frequenza di ricombinazione, che è uguale ad 1cM e corrisponde ad una lunghezza teorica di 1Mb). Dato che i geni hanno dimensioni che variano da 14kb a 2,5Mb e che la densità dei geni varia in regioni cromosomiche diverse, è impossibile stabilire a priori se una regione cromosomica mappata da marcatori genetici includa uno o più geni. Ciò è possibile solo quando la regione cromosomica mappata è stata completamente sequenziata e nella sequenza sono state individuate le sequenze dei geni.
Alcuni autori usano il termine binomiale "clonazione posizionale" in senso generale e "clonazione dei geni candidati posizionali" solo quando più geni sono presenti nella regione cromosomica dove mappa la patologia monogenica.
In genere, la mappatura dei geni dei quali è ignota la proteina codificata e la funzione fisiologica, è resa possibile quando per una mutazione essi perdono la loro funzione fisiologica e causano dei sintomi clinici.
L'interesse per la cura dei pazienti portatori di malattie genetiche ha indotto ed induce a definire lo stato di malattia sulla base di definiti sintomi, a costruire alberi genealogici delle famiglie in cui si è manifestata la malattia ed a mappare il gene responsabile della suscettibilità alla malattia. La conoscenza della posizione subcromosomica del gene ignoto responsabile della patologia monogenica permette di individuarlo mediante le strategie della clonazione

posizionale o del candidato posizionale. Nel 1986, il gene della suscettibilità alla patologia monogenica della granulomatosi cronica, è stato il primo gene identificato e clonato con la strategia della clonazione del candidato posizionale e nel 1993 con la stessa strategia erano stati isolati complessivamente 19 geni responsabili di altrettante patologie.
L'isolamento dei geni responsabili delle malattie genetiche ha permesso di individuare i geni normali della suscettibilità a quelle stesse patologie e quindi di migliorare la conoscenza della funzione cellulare persa con la patologia.
Per la clonazione posizionale è di fondamentale importanza la conoscenza delle sequenze vicine al gene, cioè di marcatori genetici del gene di interesse (sequenze RFLP, VNTR e SNP) (Tabella 3-2). Le sequenze nucleotidiche dei marcatori sono utilizzate per clonare il gene arrivando ad esso attraverso la definizione della sequenza della regione di DNA cromosomico che intercorre tra il marcatore ed il gene di interesse, camminando sul cromosoma (vedere dopo). Raggiunto il gene di interesse, viene determinata la sua sequenza nucleotidica e conoscendo la sequenza del gene è possibile mediante PCR analitica clonare il relativo cDNA al fine di dedurre da esso la sequenza aminoacidica della proteina (figure 1-14 e 1-15). Per stabilire con certezza che il gene è responsabile della patologia monogenica si analizzano gli alleli mutati dello stesso gene in portatori della stessa patologia in altre famiglie appartenenti ad aree geografiche e/o a etnie diverse. Successivamente per definire la patogenesi molecolare della malattia si utilizzano le tecnologie molecolari per individuare l'attività molecolare della proteina e la sua funzione nella fisiologia della cellula e dell'organismo e le alterazioni provocate dalle mutazioni sulla struttura e sull'attività molecolare della stessa proteina (capitolo 2).

Camminare e saltare sul cromosoma per definire la sequenza di un gene mappato geneticamente.

Le tecniche "camminare sul cromosoma" e "saltare sul cromosoma" sono utilizzate quando si è individuato un marcatore genetico del gene di interesse, del quale gene conosciamo solo il fenotipo patologico, nulla della sua sequenza e nulla della proteina codificata. Nella maggior parte dei casi studiati, il gene di interesse è divenuto tale perché ha un allele mutato responsabile di una patologia monogenica ed i sintomi della patologia sono il fenotipo che ha guidato la ricerca del marcatore genetico del gene (capitolo 3).
Queste due tecniche sono state usate per clonare nel 1989 il gene della suscettibilità alla mucoviscidosi (fibrosi cistica). Un gene di notevoli dimensioni (circa 250kb) che ha richiesto quattro anni di lavoro a ricercatori appartenenti a più laboratori. Questa apparente lentezza era causata dalla minore disponibilità, rispetto ad oggi, di marcatori genetici e di macchine per la determinazione della sequenza nucleotidica del DNA. Questi miglioramenti hanno portato a non usare più la tecnologia del saltare sul cromosoma.

Camminare sul cromosoma.
Il “camminare sul cromosoma” inizia con il vaglio di una genoteca genomica utilizzando la tecnica della PCR analitica specifica per la sequenza locus specifica del marcatore genetico del gene di interesse oppure usando come sonda la stessa sequenza locus specifica del marcatore (figura 4-2).
Tra i cloni isolati si fa la sequenza del clone più lungo avente l’estremo-3’ verso il gene da identificare. Talvolta l'orientamento della sequenza del marcatore non è noto. L'orientamento è la posizione del 5' di una sequenza di DNA rispetto al telomero del braccio p del cromosoma. L'orientamento della sequenza del marcatore è individuato insistendo nell'analisi della sequenza del DNA genomico fino ad incontrare la sequenza della quale siano note la posizione e l'orientamento sul cromosoma. Questo è necessario per capire se ci stiamo avvicinando o allontanando dal gene da individuare perché, quando si frammenta il DNA genomico, si perdono sia la posizione subcromosomica che l'orientamento dei frammenti. La determinazione della sequenza nucleotidica dei frammenti di DNA non dà indicazioni sul loro locus né sul loro orientamento nel cromosoma.
Isolato il primo clone, viene determinata la sequenza dell'inserto e stabilito il suo l'orientamento verso il gene da isolare. Dell'inserto viene preso un frammento (dalla parte, ormai nota, verso il gene da ricercare) ed usato come sonda per vagliare di nuovo la genoteca. E così in avanti, scegliendo sempre i cloni più lunghi (perché permettono il passo più lungo). I cloni isolati hanno tutti l'ultima parte dell'estremo-3’ (verso il gene ricercato) identica a quella all’estremo-5’ del clone successivo. In questo modo si può procedere verso il gene ed alla fine avere una sequenza continua e completa che lo include.
Utilizzando la tecnica della PCR, la genoteca è vagliata utilizzando 2 primer complementari alle regioni uniche nel genoma che identificano il locus del marcatore genetico di ricombinazione (es. microsatellite). Individuato e sequenziato il primo clone si procede come indicato sopra e si sintetizzano due primer che permettano l’amplificazione della regione del 3’ del primo frammento clonato e così avanti (figura 4-2).
La costruzione delle mappe di associazione dense di marcatori e delle genoteche genomiche BAC aventi cloni con lunghi inserti di DNA umano e l’avvento della tecnica della PCR dei sequenziatori automatici, hanno semplificato e reso più veloce la tecnologia del camminare sul cromosoma. La tecnica della PCR si è imposta perché può utilizzare piccole quantità di DNA stampo e per la semplicità di esecuzione. La minore distanza fisica tra marcatori genetici ha reso più breve il cammino da percorrere tra marcatore e gene di interesse, i grossi inserti delle nuove genoteche genomiche (passi più lunghi) hanno ridotto il numero dei cloni da ricercare.
Per coprire una distanza fisica di 1Mb (considerata breve) occorrono 7-8 passi con cloni BAC aventi inserti di DNA di 150.000b e, poiché un sequenziatore automatico esegue la determinazione di circa 400 basi per volta (passo del sequenziatore), occorrono circa 400 analisi consecutive (camminando sul DNA dell'inserto) per determinare la sequenza di ogni inserto e in totale circa 3000

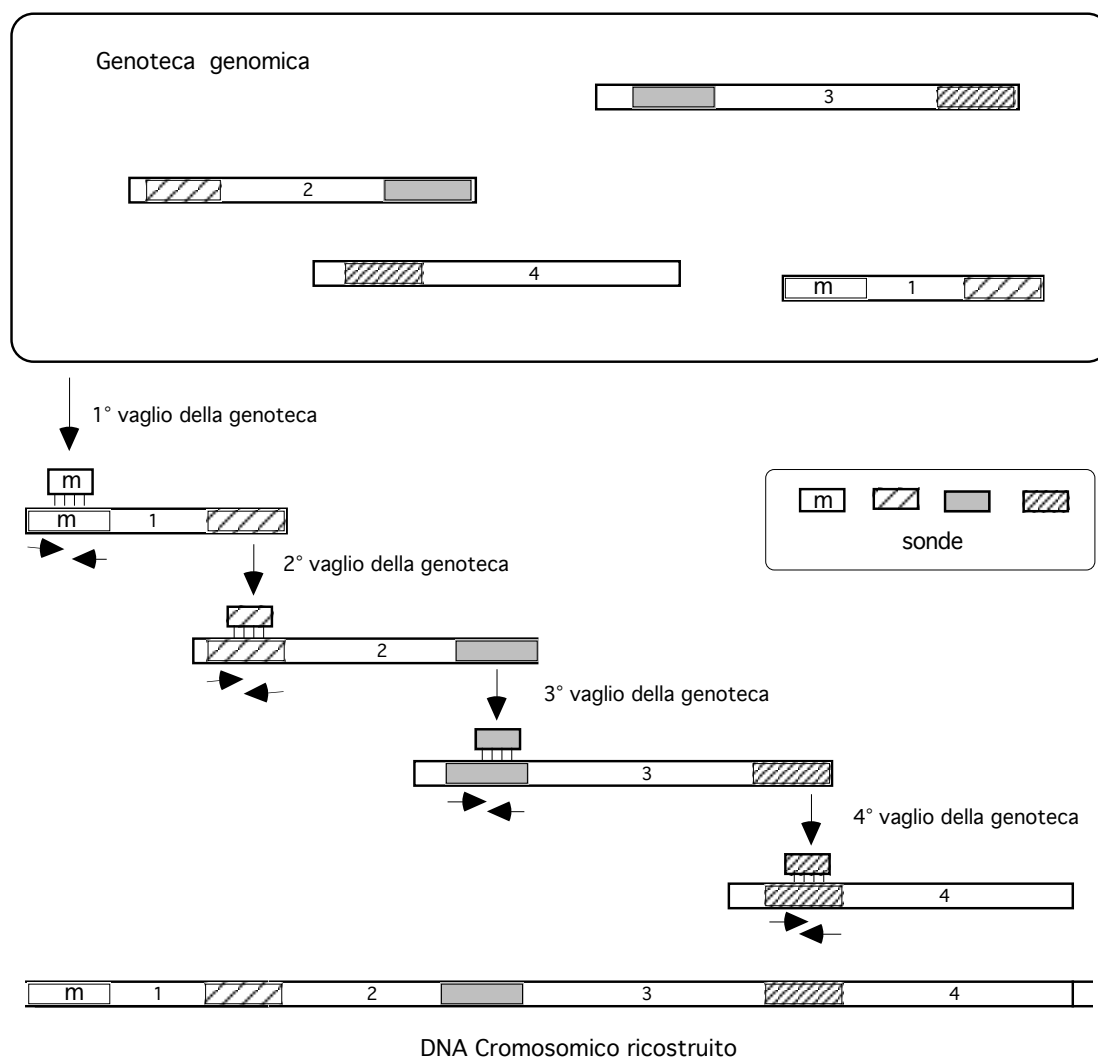



Figura 4-2. Tecnica del camminare sul cromosoma.
Per semplificare la figura la genoteca genomica è rappresentata con solo 4 inserti di DNA e non sono indicati i vettori a cui sono legati gli inserti, né le cellule contenenti i costrutti. Per altri dettagli vedere figure 1-12 , 1-13 e 1-14. Il 1° vaglio della genoteca viene fatto con la tecnica della PCR, i cui primer amplificano la sequenza, locus specifica, del marcatore del gene di interesse. I primer sono indicati con frecce disposte orizzontalmente sotto gli inserti di DNA. Il 1° vaglio porta ad isolare il primo clone che include un inserto di DNA (1° inserto). L'inserto viene liberato dal vettore, purificato e sequenziato (primo passo). La sequenza al 3' terminale dell'inserto n°1 viene utilizzata per stabilire la sequenza di due primer che mediante PCR possano amplificare la stessa regione. Con questi primer viene vagliata di nuovo la stessa genoteca genomica e vengono isolati i cloni che la contengono. Tra questi ci sarà anche il clone contenente il 1° inserto, pertanto gli inserti saranno tutti isolati ed il più lungo sequenziato (2° inserto, secondo passo). La sequenza della regione al 3' terminale del 2° inserto viene utilizzata per stabilire la sequenza di nuovi primer per vagliare di nuovo la genoteca genomica. E così avanti fino aver isolato e sequenziato tutti gli inserti della regione genomica di interesse. La parziale sovrapposizione delle sequenze dei vari inserti permette di ricostruire la sequenza del DNA della regione. In figura è indicato anche l'isolamento dei frammenti fatto mediante sonde. La prima sonda (m) è costituita da un frammento di DNA avente sequenza identica alla sequenza, locus specifica, del marcatore genetico del gene di interesse. Dal 1° inserto viene tagliata una regione all'estremo 3' (rettangolo a righe trasversali) che viene utilizzata come sonda per effettuare il 2° vaglio della genoteca. E così avanti fino a ricostruire l'intera regione del DNA cromosomico. La presenza di frammenti sovrapponibili è data da una parziale digestione del DNA genomico con uno o due enzimi di restrizione (figura 1-13). Per altri dati vedere testo (ridisegnato e modificato da Recombinant DNA,Watson J.D., Gilman M., Witkowski J. and Zoller M. (1992) 2nd ed., Scientific American Books, Freeman, USA).

analisi della sequenza per raggiungere il gene di interesse. Definita la sequenza della regione cromosomica mappata occorre individuare in essa il gene di interesse. Se non conosciamo niente del gene di interesse non è facile la presenza della sequenza di un gene esaminando una sequenza di DNA. Un metodo che viene usato consiste nel vagliare genoteche di cDNA di molti tessuti, incluso l'encefalo fetale che è l'organo che esprime il maggior numero di geni, usando come sonda il DNA del gene di interesse o suoi frammenti. Individuato il clone, il cDNA contenuto in esso indicherà che nella regione mappata è inclusa una sequenza codificante (esone) di un gene. Determinata la sequenza del cDNA, dal suo confronto con la sequenza del DNA della regione mappata si ottiene la struttura in esoni ed introni del gene di interesse.
Questo procedimento è stato semplificato dalla conservazione nelle banche dati delle sequenze EST, per cui è sufficiente fare una ibridazione *in silice* (utilizzando il programma BLAST) della sequenza del DNA della regione cromosomica mappata con le sequenze EST dell'archivio elettronico per avere l'indicazioni degli esoni del gene di interesse, purché le sequenze EST siano già state sequenziate e conservate negli archivi elettronici.

Saltare sul cromosoma.
Questa tecnica fu utilizzata per clonare il gene della patologia monogenica mucoviscidosi (fibrosi cistica) data la sua notevole lunghezza e non è stata più utilizzata in seguito al miglioramento di altre tecnologie e l'introduzione di nuove. E' qui riportata come esempio di un eccellente prodotto dell'ingegno.
La tecnica "saltare sul cromosoma" (figura 4-3) richiede la costruzione di una particolare genoteca detta dei salti che viene utilizzata quando la regione in cui mappa la patologia è molto grande (alcuni Mb), cioè quando i marcatori genetici distano molto dal locus dell'allele responsabile della patologia.
Per costruire la genoteca dei salti, il DNA genomico, ad esempio umano, venne digerito parzialmente con l'enzima di restrizione MobI in modo da poter avere anche frammenti molto lunghi. I frammenti di lunghezza tra 80-150 kb separati mediante elettroforesi dagli altri frammenti (più lunghi e più corti), furono estratti dal gel di elettroforesi e, con l'enzima ligasi, resi circolari con i due estremi legati al DNA del gene del tRNA soppressore (*supF*). Questo è un gene mutato che codifica un tRNA che riconosce un codice di stop come codice dell'aminoacido Glu ed alla traduzione introduce il glutammato nel nascente polipeptide. Successivamente con l'azione di EcoRI, enzima di restrizione diverso da quello utilizzato inizialmente per digerire il DNA umano, venne tagliata via la maggior parte del DNA umano che formava l'anello, mentre non venne tagliato il DNA *supF* perché non contiene siti EcoRI. Al termine della digestione si ottenne un costrutto costituito dal DNA del gene *supF* che ai suoi estremi ha legati due frammenti del DNA genomico umano. I due frammenti erano le regioni agli estremi 5' e 3' del frammento di DNA genomico che era stato legato al *supF* e le loro sequenze si trovavano sul DNA cromosomico separate da 80-150 kb (lunghezze dei salti sul cromosoma). Quindi fu costruita

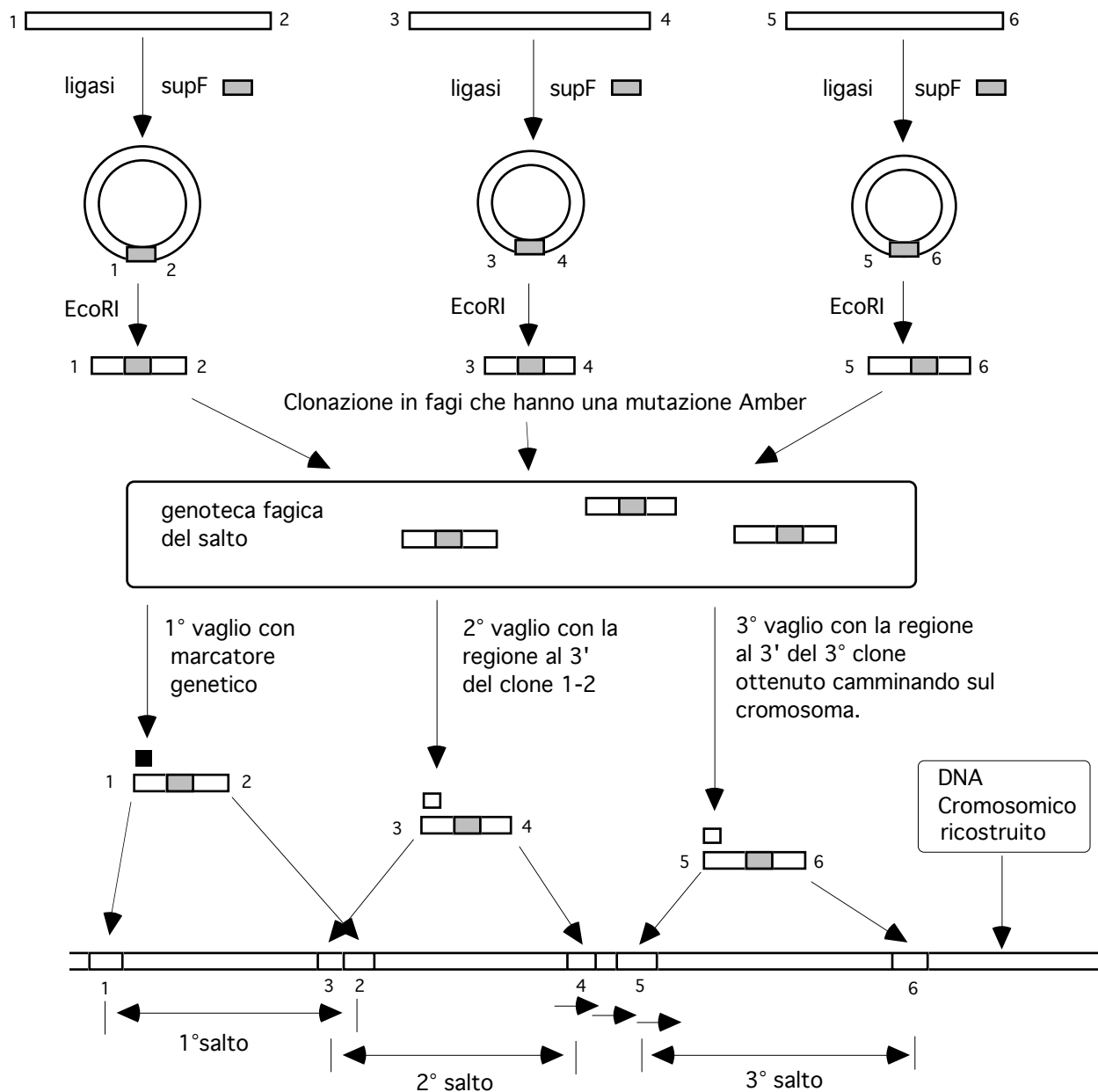



Figura 4-3. Tecnica del saltare sul cromosoma. In alto sono indicati i frammenti di DNA genomico di 80-150b che vengono legati con il DNA del gene *supF* e chiusi ad anello per azione dell'enzima ligasi. Quindi gli anelli sono tagliati con l'enzima di restrizione EcoRI che elimina gran parte del DNA genomico. I frammenti digeriti con EcoRI, che risultano avere una parte centrale (gene *supF*) avente ai lati frammenti di DNA genomico, sono clonati in fago. La genoteca fagica del salto è poi vagliata con un marcatore della regione cromosomica e viene isolato il clone positivo il cui inserto è il frammento 1-2. Sul cromosoma la sequenza dell'estremo 3' (n°2) del frammento 1-2 risulterà essere distante 80-150b dalla sequenza dell'estremo 5' (n° 1) dello stesso clone. Parte dell'estremo 3' del frammento 1-2 è utilizzata per vagliare nuovamente la genoteca dei salti e porta all'isolamento del frammento 3-4. Può accadere che l'estremo al 3' di un frammento non includa regioni del frammento più vicino (salto corto) per cui la regione mancante viene completata con il camminare sul cromosoma (in figura le tre frecce orizzontali tra 4 e 5). Il terzo passo, utilizzato come sonda, permette di isolare dalla genoteca dei salti il frammento 3-6. In questo modo si può ricostruire la regione cromosomica da 1 a 6. Per altri dettagli vedere testo. (ridisegnato e modificato da Recombinant DNA, Watson J.D., Gilman M., Witkowski J. and Zoller M. (1992) 2nd ed., Scientific American Books, Freeman, USA).

la genoteca dei salti, i costrutti [DNA umano]-[DNA-*supF*]-[DNA umano] furono legati al DNA di fagi che portano una mutazione amber (mutazione del codice dell'aminoacido GLU che lo converte in codice stop). Con i fagi della genoteca dei salti furono infettati batteri anch'essi mancanti del gene *supF*, cioè incapaci di replicare se non quando vengono infettati da fagi che portano il costrutto contenente *supF* che sopprime la mutazione amber dei batteri. In questo modo replicheranno solo i fagi che hanno inserito il costrutto. Si formano così le placche di lisi nelle colonie batteriche contenenti i costrutti con gli inserti di interesse. Le placche dei batteri lisati inizialmente furono vagliate (figura 1-14) con sonde di DNA del marcatore della regione contenente il gene da ricercare. Furono isolati dei cloni aventi all'estremo-5' la sequenza identica a quella della sonda ed all'estremo-3' la sequenza posta a 80-150kb lontano dalla prima nella direzione del gene da isolare (1° salto in figura 4-3). Parte dell'estremo-3' venne tagliata ed utilizzata come sonda per isolare altri cloni dalla genoteca dei salti, e così avanti. In figura 4-3 è presentato il caso in cui con la sonda costituita dalla sequenza-4 (regione all'estremo-3' del frammento 3-4) non fu possibile isolare cloni dalla genoteca del salto, perché non c'era sovrapposizione di sequenza tra i frammenti 3-4 e 5-6. Tuttavia la stessa sonda-4 fu utilizzata per vagliare una normale genoteca genomica e quindi per procedere camminando sul cromosoma. Ogni volta che camminando sul cromosoma veniva isolato un clone, esso (o parte di esso al 3') era utilizzato come sonda per vagliare la genoteca dei salti con lo scopo di trovare un nuovo clone e fare un altro salto sul cromosoma. In figura 4-3 con il clone del 3° passo viene isolato il clone del 3° salto.
Questa tecnologia permette di compiere salti lunghi fino a 150kb equivalenti a più passi della tecnica del camminare sul cromosoma. In questo modo viene rapidamente mappata fisicamente una vasta regione cromosomica e muovendosi dalle sequenze dei salti si può procedere a clonare e sequenziare le regioni cromosomiche vicine o all'interno degli stessi salti.

Strategia della clonazione dei geni candidati posizionali

I sintomi della patologia monogenica di interesse mappano in una regione relativamente grande perché i marcatori sono distanti dal locus della patologia più di un 1cM e la regione cromosomica mappata include più geni da considerare tutti candidati posizionali.
La strategia della clonazione del candidato posizionale è sempre un po' mista perché qualche sintomo della patologia deve essere noto (altrimenti la patologia che patologia è), tuttavia i sintomi non danno indicazioni sulle funzioni alterate tali da poter proporre qualche proteina nota come candidata funzionale ed utilizzare su essa la strategia della clonazione funzionale dei geni (es. patologia AME, Apparent Mineralcorticoid Excess, insorgenza del tumore della mammella).

Utilizzando come sonda un marcatore genetico della regione si comincia a clonare ed analizzare la sequenza della regione di cromosoma includente due o più geni (vedere figura 4-2).
In relazione alla precisione della mappatura del gene della patologia (distanza genetica tra marcatore e gene) e quindi alla dimensione della regione cromosomica analizzata possono essere presenti anche decine di geni candidati posizionali.
Nel caso in cui più di un gene clonato risulti aver un allele associato stretto alla patologia monogenica, tutti quegli alleli sono candidati posizionali ad essere responsabili della patologia ed i relativi geni sono tutti candidati posizionali ad essere responsabili della suscettibilità alla patologia monogenica. Il gene della suscettibilità alla patologia è individuato quando si dimostra che è l'unico gene, tra tutti quelli candidati, ad avere un allele portatore di una mutazione distruttiva e/o essere presente nei membri malati e non nei membri sani di numerose famiglie appartenenti a regioni geografiche e/o etnie diverse.
L'allele mutato è definito responsabile della patologia monogenica anche se non si conoscono l'attività molecolare e la funzione cellulare della proteina da esso codificata.
La conoscenza delle caratteristiche molecolari e funzionali della proteina normale e mutata sono utili per confermare la responsabilità delle mutazioni osservate sul gene ed inoltre per avere informazioni sulla patogenesi molecolare, cioè sui meccanismi molecolari che la proteina alterata attua causando i sintomi e le alterazioni che caratterizzano la patologia. Tuttavia questa ricerca può richiedere molti anni di lavoro sperimentale (vedere dopo la ricerca del gene BRCA1).
Quanto detto sopra circa le strategie di clonazione funzionale e posizionale rende evidente che la clonazione di molti geni è stata possibile perché sono state utilizzate le conoscenze accumulate in molti anni di ricerche sulla sequenza, attività molecolare e funzione cellulare delle proteine e sulla mappatura di molte patologie monogeniche.

Progetto genoma umano

Il progetto genoma umano, proposto ed attuato principalmente per la volontà del premio Nobel James Watson, fu iniziato nel 1995. Il progetto fu finanziato per 10 anni di lavoro e occupò molti ricercatori appartenenti a laboratori sparsi in tutto il mondo che si erano divisi i compiti, cioè i cromosomi. Il miglioramento delle tecnologie del DNA, avvenuto durante quegli anni, permise di concludere il progetto nel 2003, con 2 anni di anticipo.
La sequenza completa del DNA dei cromosomi mitocondriali umani data la sua piccola dimensione (16.599 basi) era stata completata nel 1981.
La determinazione della sequenza del DNA dei cromosomi aveva lo scopo di individuare tutti i geni umani. La sua importanza è intuibile considerando che

negli anni '90 era nota, come struttura e/o come attività molecolare, molto meno della metà delle proteine umane ed appariva insormontabile identificare e purificare proteine presenti nelle cellule in bassissime concentrazioni, così come quelle espresse solo in piccoli gruppi di cellule di un dato organo (es. nuclei del sistema nervoso) o espresse in brevi fasi dello sviluppo embrionale. Inoltre di molte proteine non si conosceva nemmeno l'esistenza. Egualmente era difficile mappare alcune patologie soprattutto quelle poligeniche.
Non avendo informazioni sulla struttura o sull'attività della proteina, normale o patologica, l'unico modo per identificare i geni ancora ignoti era quello di determinare la sequenza delle molecole di DNA di tutti i cromosomi.
La difficoltà ad individuare le proteine presenti in un essere vivente con metodi esclusivamente biochimici è stata dimostrata dal completamento dell'analisi del genoma del Escherichia coli, il batterio più studiato a livello biochimico e del quale si conosceva la funzione di 1800 geni. L'analisi totale del DNA del suo genoma ha stupefatto molti esperti microbiologi perché ha mostrato che lo Escherichia coli ha ben 4288 geni codificanti proteine. Di ben 2488 geni non si era nemmeno ipotizzata la presenza.
Si assume che i nuovi geni, individuati con il progetto genoma umano, se mutati possano essere responsabili della suscettibilità a patologie delle quali sono noti i sintomi ma non le cause genetiche.
Le strategie funzionali e posizionali sono tuttora utilizzate in forma più semplificata (vedere dopo gli schemi delle clonazioni nel 2005) perché il completamento della sequenza del DNA di tutti i cromosomi umani e la conservazione delle sequenze nelle banche dati ha eliminato la necessità di mappare e clonare i geni. Tuttavia, dopo aver determinato la sequenza di un DNA proveniente da nuovi individui sani o malati, è sempre necessario confrontarla con le sequenze conservate nelle banche dati, sopratutto per la possibile presenza del polimorfismo e per possibili errori di sequenza che sono calcolati intorno al 1% (1 ogni 100 basi).

Schemi delle strategie per la clonazione dei geni

Strategia della clonazione funzionale dei geni.

Schema della strategia della clonazione funzionale dei geni:
Analisi specifica della proteina → clonazione e sequenza del cDNA → clonazione e sequenza del gene → mappatura fisica del gene.

La strategia della clonazione funzionale permette di clonare e mappare fisicamente un gene di interesse quando si conosca un metodo di analisi specifico per la molecola o per l'attività molecolare della proteina da esso codificata.

1. Per procedere alla clonazione del gene occorre avere almeno uno dei seguenti dati sperimentali:

a) la conoscenza della sequenza aminoacidica della proteina (o parziale di almeno 7 aminoacidi). Conoscendo la sequenza della proteina, si conosce anche il tessuto da cui è stata purificata. Dal tessuto si estrae lo mRNA totale, si converte in cDNA e si costruisce una genoteca di cDNA che è vagliata utilizzando come sonda un insieme di oligonucleotidi degenerati, la sequenza dei quali è dedotta dalla sequenza aminoacidica della proteina, tenendo conto della degenerazione del codice genetico.
b) la conoscenza del dosaggio dell'attività molecolare della proteina. Si costruisce una genoteca di espressione di cDNA e si vaglia utilizzando il dosaggio dell'attività molecolare della proteina.
c) possesso dell'anticorpo specifico contro la proteina. Si costruisce una genoteca di espressione di cDNA e si vaglia utilizzando l'anticorpo.
d) il possesso di cellule coltivate mutate che mancano dell'attività molecolare della proteina codificata dal gene di interesse. Si utilizzano le cellule come riceventi del cDNA di cellule normali per costruire la genoteca e mediante il saggio di complementazione, si isola il clone contenente il cDNA che esprime la proteina di interesse.
Inizialmente si clona il cDNA e non direttamente il gene di interesse perché la costruzione ed il vaglio delle genoteche di cDNA è più semplice (meno cloni) di quelle genomiche. Inoltre quando non è nota la sequenza della proteina occorre operare con genoteche di espressione di cDNA dato che è praticamente impossibile costruire genoteche genomiche di espressione. Inoltre la digestione del DNA genomico con enzimi di restrizione può casualmente dividere i geni in due o più frammenti rendendoli biologicamente inattivi.
2. Isolato il cDNA di interesse se ne determina la sequenza.
3. Si costruisce una genoteca genomica e si clona il gene di interesse vagliando la genoteca genomica mediante una PCR specifica per il cDNA o utilizzando il cDNA (o parte di esso) come sonda.
4. Il gene è mappato fisicamente vagliando la mappa delle sequenze STS, utilizzando come sonda la sequenza al 3'UTR del gene. I loci di molti marcatori fisici STS sono mappati anche sulle mappe citogenetica e genetica, pertanto se il gene di interesse mappa su una banda citogenetica o su una regione della mappa genetica sulla quale mappa una patologia monogenica esso diviene candidato posizionale della suscettibilità a quella patologia.
5. La proteina codificata dal gene può essere espressa *in vitro* mediante transfezione per ricercare o per migliorare la conoscenza delle sue caratteristiche strutturali, la sua attività molecolare e la sua funzione cellulare (capitolo 2).

Strategia della clonazione funzionale dei geni nell'anno 2005

Schema della strategia della clonazione funzionale dei geni nell'anno 2005:
clonazione funzionale della proteina → sequenza del cDNA → ricerca elettronica della sequenza del gene→ definizione elettronica del locus citogenetico, genetico e fisico del gene.

Alcune fasi della strategia funzionale sono rese più semplici dalla possibilità di analizzare elettronicamente nelle banche dati (es. www.ncbi.nlm.nih.gov), utilizzando il proprio computer, la mappa della sequenza del DNA del genoma umano nucleare e mitocondriale, le mappe genetiche, in particolare quelle costituite da microsatelliti e SNP e gli altri archivi elettronici contenenti i dati dei geni umani: sequenza del gene, del relativo cDNA e delle EST, sequenza e caratteristiche della proteina codificata (attività molecolare, funzione cellulare, localizzazione tissulare e subcellulare, ecc.). La sequenza ed il locus della quasi totalità dei geni sono noti, mentre molte (circa 2/3) delle proteine da essi codificate sono tuttora ignote (la proteina non è stata purificata, non si conosce la sua funzione cellulare e/o l'attività molecolare) e la loro esistenza è dedotta dalla presenza della sequenza del gene nella mappa della sequenza del DNA genomico.

1. Quando si sia determinata la sequenza aminoacidica di un peptide di almeno 7 aminoacidi della proteina di interesse (il peptide è codificato da una sequenza di 21b) non occorre più clonare il gene perché mediante un particolare programma di gestione (**tblastn**) si può individuare il gene ricercando in una banca dati della sequenza del genoma umano (es. http://www.ncbi.nlm.nih.gov) la sequenza nucleotidica che codifica il peptide della proteina di interesse. Il programma analizza le sequenze nucleotidiche di EST, cDNA e dei geni conservate nella banca dati, traducendole in sequenze aminoacidiche e confronta queste sequenze con quella del peptide di interesse. Al termine dell'operazione (in genere dopo pochi secondi o pochi minuti) le sequenze aminoacidiche includenti quella del peptide di interesse saranno allineate in ordine decrescente del percento di similarità (figura 1-18).

Se la sequenza del peptide è identica ad una tradotta da una sequenza del DNA genomico si ha il risultato importante dell'identificazione del gene di interesse, ma non la prova definitiva dell'identificazione stessa, perché la sequenza del peptide potrebbe appartenere ad una proteina simile a quella trovata nella banca dati. Le due proteine potrebbero avere identica solo la sequenza del peptide utilizzato per ricercare elettronicamente il gene di interesse, ma non la rimanente parte della sequenza. Al fine di conoscere l'intera sequenza aminoacidica della proteina di interesse viene clonato il relativo cDNA, mediante la clonazione funzionale e utilizzando gli oligonucleotidi degenerati. Nel caso che la sequenza della proteina di interesse sia completamente ignota si clona il cDNA con la clonazione funzionale utilizzando il dosaggio dell'attività molecolare della proteina, con gli anticorpi anti-proteina di interesse o con il saggio di complementazione funzionale. Quindi si usa un programma di gestione (**tblastx**) che traduce la sequenza del cDNA in quella della proteina e la confronta con le sequenze aminoacidiche tradotte dal DNA di tutti i cromosomi umani e le sequenze allineate sono mostrate in ordine decrescente del percento di similarità. Al primo posto ci può essere una proteina avente la sequenza identica a quella della proteina di interesse, cioè un cDNA conservato nella banca dati codifica esattamente la proteina di interesse. In questo caso la

proteina di interesse è codificata da un cDNA e quindi da un gene già individuato. Dalla pagina elettronica del cDNA, con un click, si può attivare il codice del relativo gene e passare alla pagina che fornisce tutti i dati del gene: la sequenza degli esoni ed introni, la connessione elettronica per la pagina della proteina codificata ed altri dati. Può accadere che il polipeptide codificato dal cDNA di interesse risulti allineato con un polipeptide diverso per uno o pochi aminoacidi. In questo caso la proteina di interesse è codificata da una variante allelica, non ancora nota, del gene conservato nella banca dati, oppure la proteina di interesse è codificata da un gene duplicato da quello conservato nella banca dati e non ancora identificato come duplicato. Utilizzando il programma BLAT e la banca dati che include la mappa della sequenza del DNA (http://genome.ucsc.edu) si mappa fisicamente il cDNA di interesse. Dalla posizione subcromosomica fisica del gene di interesse si ha l'indicazione che esso sia una variante allelica (stesso locus fisico) o un gene duplicato (locus fisico diverso) del gene conservato nella banca dati.
Mediante il programma di gestione align sono confrontate mediante allineamento le sequenze nucleotidiche del cDNA di interesse con quelle individuate nelle banche dati.
L’allineamento dei due cDNA mostrerà:
- l’identità delle due sequenze; si assume che siano codificate dallo stesso allele dello stesso gene.
- la diversità delle sequenze per mutazioni silenti o per mutazioni conservative; si assume che siano codificate da alleli diversi di uno stesso gene o di un gene ripetuto.
- la diversità per estese parti della sequenza; si assume che siano codificate da geni diversi.
Ogni proteina conservata nelle banche dati (e così il cDNA ed il gene) ha un proprio codice di riconoscimento attivabile elettronicamente che permette di collegarsi alla pagina che include i suoi dati (sequenza, specie e tessuto di origine, pubblicazioni scientifiche che documentano i dati). Dalla stessa pagina elettronica si può passare direttamente ad altri archivi che forniscono la sequenza del relativo cDNA e del gene, la posizione cromosomica e subcromosomica del gene nella mappa fisica, nella mappa genetica ed in quella citogenetica. Egualmente dalla pagina elettronica del gene si può passare a quella della proteina e a quella del cDNA o dalla pagina del cDNA procedere alle altre pagine.
Il gene, individuato elettronicamente, può essere un gene già noto e può essere nota anche la proteina da esso codificata, in questo caso la ricerca del gene codificante la proteina di interesse è conclusa. Oppure, il gene individuato elettronicamente può essere stato individuato nella sequenza del DNA nucleare mediante ibridazione elettronica con le EST (delle quali ancora non si conosce la proteina codificata) e/o sulla base delle sequenze consenso comuni ad altri geni (sequenze promotrici, sequenze segnale di splicing, sequenza segnale di poliadenilazione). In questo caso, non si conosce la funzione del gene di interesse e, mediante la clonazione funzionale e le analisi elettroniche delle

sequenze geniche conservate nelle banche dati, si riesce ad attribuire al gene la sua funzione dimostrando che codifica una proteina di nota attività molecolare e funzione cellulare.
Se dalla ricerca elettronica si ha l'indicazione che la proteina di interesse è codificata da una variante genetica del gene conservato nelle banche dati, si può sintetizzare *in vitro* la proteina mediante transfezione del cDNA, analizzare la sua attività molecolare (capitolo 2) e confrontarla con quella della variante nota conservata nell'archivio elettronico.

Strategia della clonazione dei geni candidati funzionali

Si ipotizza che una proteina della quale sono note l'attività molecolare e la funzione cellulare, se alterata geneticamente, causi le stesse alterazioni di una patologia monogenica. Il gene che la codifica è il gene candidato funzionale.

Schema della strategia della clonazione dei geni candidati funzionali:
Proteina candidata funzionale → clonazione funzionale del gene candidato → ricerca ed identificazione dei marcatori genetici del gene → associazione genetica di un allele del gene candidato alla patologia → dimostrazione che l'allele associato alla patologia è anche mutato → dosaggio dell'attività molecolare della proteina mutata sintetizzata *in vitro* mediante transfezione → applicazione della stessa strategia in altre famiglie.

La strategia del gene candidato funzionale permette di clonare un gene e di identificarlo come gene della suscettibilità ad una patologia genetica monogenica, assumendo e poi dimostrando che la patologia sia causata da una mutazione del gene che altera l'attività molecolare e la funzione cellulare della proteina codificata.

1. La strategia può essere applicata quando si conosce l'attività molecolare (es. attività di catalisi) e la funzione cellulare (es. enzima di una via metabolica di sintesi di un ormone steroideo, o proteina di membrana plasmatica deputata al trasporto di metaboliti all'interno della cellula) di una proteina e si ipotizza che un'alterazione della sua attività molecolare e quindi della sua funzione cellulare causi le stesse alterazioni e quindi gli stessi sintomi causati da una patologia monogenetica nota.
Il gene che codifica la proteina candidata, è ipotizzato essere il gene candidato alla suscettibilità a quella patologia, cioè se mutato causa la patologia.
2. Se il gene che codifica la proteina candidata funzionale (gene candidato funzionale) è già stato clonato si passa direttamente al punto 3.
Se il gene candidato non è stato ancora clonato si clona con la strategia funzionale. Con la clonazione funzionale si ottiene il cDNA della proteina candidata ed utilizzando la sequenza del cDNA si sintetizzano dei primer per effettuare una PCR specifica per il cDNA e quindi anche per il relativo gene. Con la PCR si vaglia una genoteca genomica, si clona e sequenzia il gene intero e, con il confronto delle sequenze del cDNA e del gene, si stabilisce la struttura in

esoni-introni del gene stesso (la conoscenza della sequenza degli esoni è utilizzata dopo al punto 5).
3. Ricerca ed identificazione del/i marcatore genetico del gene candidato.
Un metodo consiste nel camminare sul cromosoma ricercando, al 5' e 3' del gene ed all'interno degli introni, dei marcatori genetici polimorfici. Si preferisce utilizzare i microsatelliti, per il loro alto polimorfismo, la loro distribuzione su molti loci, per la possibilità di analizzarli con la tecnica PCR analitica e perché è più facile individuarli leggendo la sequenza del DNA (si notano le sequenze ripetute analizzando le sequenze di un singolo allele) rispetto ai marcatori RFLP o SNP (per individuarli occorre confrontare le sequenze dei due alleli). I marcatori genetici all'interno degli introni sono ovviamente associati stretti al gene e sono assunti essere associati stretti anche i microsatelliti distanti circa 100kb dal 5' o 3' dal gene candidato. Il grado di associazione tra marcatore e gene è verificato facendo l'analisi della associazione degli alleli di ogni marcatore con gli alleli del gene candidato nei membri delle famiglie con alta incidenza della patologia.
Si preferisce fare l'analisi dell'associazione della patologia con i marcatori del gene piuttosto che con il gene stesso perché l'analisi del marcatore è fatta con una unica PCR, mentre dovendo analizzare il gene occorrerebbero molte più analisi per verificare la presenza di polimorfismo all'interno di ciascun esone (vedere dopo punto 6 e patologia MODY).
4. Si ricerca l'associazione di un allele del marcatore (che è associato stretto ad un dato allele del gene candidato) con la patologia (cioè con l'allele mutato del gene della suscettibilità alla patologia ancora ignoto). Si verifica questa associazione in circa 100 membri malati, generati da gameti provenienti da almeno 100 meiosi informative (capitolo 3) e si accetta che al massimo si sia verificata una sola ricombinazione tra marcatore e patologia. 1 sola ricombinazione su 100 meiosi informative corrisponde ad una frequenza di ricombinazione del 1% e ad una distanza genetica di 1cM. 1cM corrisponde mediamente ad 1Mb e questa è la distanza fisica teorica tra il marcatore del gene candidato e il gene responsabile della patologia. L'esperienza ha indicato che con queste distanze (genetica e fisica) relativamente brevi, il marcatore è associato stretto al gene candidato ed anche al gene responsabile della patologia. È considerato tecnicamente breve anche il cammino sul cromosoma da fare per percorrere la distanza tra marcatore e gene responsabile della patologia.
La frazione di ricombinazione è determinata dividendo il numero dei malati ricombinanti per il numero totale dei malati nati da meiosi informative e controllando la validità della frazione di ricombinazione mediante il calcolo del Lod-score. Quando si osservano ricombinazioni tra l'allele del marcatore genetico del gene candidato e l'allele (ancora ignoto) responsabile ad esempio di una patologia autosomica dominante, vuol dire che nella gametogenesi del genitore portatore della patologia durante la meiosi, uno dei due alleli è finito sul cromosoma omologo. Ciò può avvenire in due modi. Nel genitore malato, l'allele del marcatore genetico e quello della patologia erano sullo stesso cromosoma,

mentre nel cromosoma ricombinato trasmesso ad un suo figlio è rimasto l'allele del marcatore e l'allele responsabile della patologia è passato sul cromosoma omologo. Fortunatamente il figlio è sano (sebbene porti un cromosoma marcato come patologico) perché ha ricevuto l'allele sano del gene responsabile della suscettibilità alla patologia dominante. Nel caso che il marcatore sia passato sul cromosoma omologo (non trasmesso al figlio) e l'allele responsabile della patologia sia rimasto sulla parte di cromosoma trasmesso al figlio, il figlio è malato sebbene sia marcato da un allele del marcatore che lo indica come sano. L'analisi dell'associazione genetica tra marcatore e malattia rivela quale delle due possibilità si è verificata, e se la ricombinazione si verifica una sola volta su 100 individui generati da gameti provenienti da meiosi informative si può procedere comunque con la ricerca dell'allele responsabile della patologia.
L'analisi della frequenza di ricombinazione sopra indicata, verifica anche se il marcatore genetico sia informativo (diagnostico della patologia) in tutte le famiglie analizzate. Può accadere che un marcatore genetico sia informativo in una famiglia e non in un'altra, oppure che un microsatellite che era informativo negli ascendenti del marito, in relazione alla genetica del microsatellite di sua moglie, possa divenire non informativo nei suoi figli (figura 3-7).
Se nessun allele del gene di interesse è associato stretto alla patologia, il gene è escluso dalla candidatura.
5. Occorre verificare che l'allele associato alla patologia abbia la sequenza nucleotidica diversa da quella dell'allele/i presente nei membri sani della stessa famiglia. Per la verifica è stata molto usata la tecnica SSCP (Single-Strand Conformation Polymorphism, Polimorfismo della conformazione dei singoli filamenti, figura 4-5). Con questa tecnica si analizza la differente migrazione elettroforetica delle strutture tridimensionali che i singoli filamenti di DNAss assumono spontaneamente e che sono sequenza specifiche. Con la tecnica SSCP si possono analizzare amplificati di circa 300bp per cui in genere viene amplificato ed analizzato un intero esone; quando l'esone è più lungo di 300bp si fanno due o più amplificati con sequenze in parte sovrapposte includenti l'intero esone. Con la tecnica SSCP sono analizzati tutti gli esoni del gene candidato. La tecnica SSCP permette di poter distinguere due filamenti di DNAds diversi anche per una singola base, presenti nella stessa soluzione. Amplificando con la tecnica PCR il DNA di cellule umane diploidi, si amplifica contemporaneamente il DNAds dei due alleli (i primer si associano alle sequenze omologhe ed identiche dei due alleli) e la tecnica SSCP permette di individuare se sono dimorfici. Per individuare la posizione e la base mutata, segnalata dalla tecnica SSCP, si sequenziano gli amplificati-PCR degli esoni risultati dimorfici e si osserva la presenza di due basi diverse nella stessa posizione della sequenza (figure 1-10b e 3-10). Dato che l'automazione della tecnica per la determinazione della sequenza del DNA ha portato ad una grossa riduzione dei tempi e dei costi delle analisi, attualmente per individuare le mutazioni negli alleli di un gene si preferisce fare direttamente l'analisi della sequenza degli amplificati degli esoni senza eseguire l'analisi SSCP.

Quando la tecnica per determinare la sequenza non era ancora automatizzata si clonavano gli amplificati-PCR dimorfici (si purificano così i singoli filamenti omologhi di DNAds) e si sequenziavano i frammenti clonati.
Se la patologia è autosomica dominante, l'allele associato alla patologia deve risultare in eterozigosi nei malati ed assente nei sani, mentre se la patologia è autosomica recessiva l'allele associato alla patologia deve essere in omozigosi nei malati e assente o in eterozigosi nei sani della stessa famiglia. Le patologie che mappano sul cromosoma X sono recessive nelle femmine e dominanti nei maschi, mentre quelle dominanti sono tali sia nei maschi che nelle femmine.
La presenza nei portatori della patologia monogenica di un allele avente sequenza diversa rispetto a quella del suo omologo presente nei sani è un risultato importante, ma non una prova dell'identificazione del gene responsabile della patologia, perché l'allele variante potrebbe portare una mutazione conservativa non patologica.
6. Si sintetizza, mediante la tecnica PCR, il cDNA dell'allele (del gene candidato) associato ai membri sani della famiglia e mediante mutazione sito-specifica il cDNA dell'allele associato ai membri della stessa famiglia portatori della patologia. I cDNA sono transfettati in cellule competenti per sintetizzare le relative proteine che sono purificate e la loro attività molecolare dosata. L'attività molecolare della proteina codificata dall'allele associato alla patologia deve risultare alterata rispetto a quella della proteina codificata dall'allele/i presente/i nei membri sani della stessa famiglia. In genere l'attività molecolare della proteina mutata risulta essere nulla o inferiore a quella della proteina normale, mentre in un minor numero di patologie è incrementata (es. alcune oncoproteine mutate, appendice E).
Se si verifica ciò, l'allele mutato è definito responsabile della patologia ed il gene è definito della suscettibilità della patologia (vedere dopo la clonazione del gene della suscettibilità alla patologia MODY).
L'analisi dell'attività molecolare delle proteine mutate sintetizzate *in vitro* può contribuire a chiarire la patogenesi molecolare della malattia. Ad esempio, se la patologia è dominante ed è sufficiente la perdita di una piccola frazione dell'attività molecolare per causare un'alterazione della funzione cellulare della proteina, si ha l'indicazione che la proteina possa far parte di un meccanismo di regolazione (vedere dopo patologia MODY e appendici D ed E).
7. Ulteriori conferme dell'identificazione del gene clonato come gene della suscettibilità alla patologia monogenica si ottengono ripetendo dal punto 5 questa stessa strategia in altre famiglie, con membri portatori della stessa patologia monogenica, ogni volta che ne vengono individuate di nuove.
Per individuare le mutazioni si determina la sequenza degli amplificati con PCR eseguite con gli stessi primer usati per individuare il gene responsabile della suscettibilità alla patologia o si esegue la tecnologia SSCP (vedere punto 5 di questa strategia).
Questa ricerca ha anche lo scopo di migliorare la conoscenza dell'eziologia e patogenesi molecolare della patologia monogenica. Può risultare che gli stessi sintomi siano provocati da alleli dello stesso gene mutati diversamente (basi

diverse nella stessa posizione o mutazioni in posizioni diverse della stessa sequenza che causano la sostituzione di aminoacidi diversi nella proteina codificata). In questo caso si ha l'informazione che la patologia monogenica è dotata di eterogeneità allelica. La maggior parte delle patologie monogeniche umane sono caratterizzate da eterogeneità allelica. Può risultare che in alcune famiglie affette dalla stessa patologia, l'allele identificato come responsabile della patologia sia integro e che responsabile della stessa patologia monogenica sia l'allele mutato di un altro gene. In questo modo si dimostra che la patologia è dotata di eterogeneità genica (eterogeneità di locus) e che è provocata da cause diverse (proteine diverse mutate) e quindi potenzialmente sensibili a farmaci diversi (vedere dopo patologia MODY2 e appendice E).

Strategia della clonazione dei geni candidati funzionali nell'anno 2005

Si ipotizza che una proteina della quale sono note l'attività molecolare e la funzione cellulare, se alterata geneticamente, provochi le stesse alterazioni di una patologia monogenica. Il gene che la codifica è il gene candidato funzionale.

Schema della strategia della clonazione dei geni candidati funzionali nell'anno 2005:
proteina candidata funzionale → determinazione della sequenza di un suo peptide o della sequenza del cDNA che la codifica → ricerca elettronica della sequenza del gene e dei suoi marcatori genetici → associazione di un allele del gene candidato alla patologia → dimostrazione che l'allele associato alla patologia è anche mutato → dosaggio delle proteine mutate sintetizzate *in vitro* → applicazione della stessa strategia in altre famiglie.
Alcune fasi della strategia della clonazione dei geni candidati funzionali sono rese più semplici dalla possibilità di analizzare elettronicamente dal proprio computer le sequenze dei geni, dei cDNA e delle proteine, le mappe di associazione e fisiche ed altri dati del genoma umano che sono conservati nelle banche dati (vedere la strategia della clonazione funzionale dei geni 2005).

1. Questa strategia è proposta ogni volta che viene individuata o migliorata la conoscenza dell'attività molecolare e/o della funzione cellulare di una proteina. La modalità della candidatura del gene è identica a quella che si faceva prima del 2005: si ipotizza che l'alterazione dell'attività molecolare della proteina porti ad una alterazione della funzione cellulare identica a quella di una patologia monogenica nota.
Il gene che codifica la proteina candidata viene ipotizzato essere il gene candidato della suscettibilità ad una patologia monogenica nota (se mutato il gene causa la patologia).
Se la sequenza del gene non è ancora nota, determinando parzialmente la sequenza aminoacidica della proteina (sequenza di un peptide) si può individuare il gene che la codifica, ricercandolo nella banca dati delle sequenze genomiche umane con il programma di gestione **tblastn** che analizza e traduce le sequenze nucleotidiche del DNA in sequenze aminoacidiche. Se la proteina candidata non è

pura non si può determinare la sequenza di un suo peptide, tuttavia se si possiede un anticorpo contro la proteina o si conosce il dosaggio della sua attività molecolare si può clonare il cDNA che la codifica e poi sequenziarlo. Utilizzando la sequenza del cDNA ed il programma **blastn** si individua il gene candidato nella stessa banca dati sopra indicata.
Tramite collegamenti elettronici ad altri archivi si ottengono i dati del gene: sequenza, locus fisico, genetico e citogenetico, cDNA, sequenza della proteina da esso codificata (vedere punto 1 della strategia della clonazione funzionale di un gene nell'anno 2005).
2. Si ricercano elettronicamente i marcatori genetici polimorfici del gene candidato funzionale scorrendo gli archivi che conservano i dati della mappa genetica del genoma umano. Si possono individuare marcatori all'interno degli introni del gene (che sono sicuramente associati al gene) e marcatori genetici distanti meno di 1cM dal 5' e dal 3' del gene candidato. Si assume che marcatori distanti 1cM dal gene siano associati stretti al gene candidato, comunque la loro frequenza di associazione con il gene sarà verificata con tecnologie genetico-molecolari.
3. Si determina la frequenza di ricombinazione tra un allele del marcatore genetico e la patologia con tecnologie genetico-molecolari (vedere punto 4 della strategia dei geni candidati funzionali).
Se nessun allele del gene candidato è associato stretto alla patologia, il gene è escluso dalla candidatura.
4. Si verifica che l'allele associato alla patologia abbia la sequenza nucleotidica di uno dei suoi esoni diversa da quella dell'allele/i presente/i nei membri sani della stessa famiglia. La verifica è effettuata mediante determinazione della sequenza degli amplificati PCR degli esoni (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
Le caratteristiche genetiche dell'allele associato alla patologia devono essere coerenti con quelle della patologia monogenica di interesse (vedere punto 6 della strategia dei geni candidati funzionali).
5. Si sintetizzano mediante transfezione dei cDNA, normali e dei cDNA mutati *in vitro*, le rispettive proteine. L'attività della proteina codificata dall'allele associato alla patologia deve risultare alterata: carente o eccessiva (vedere punto 6 della strategia della clonazione dei geni candidati funzionali).
Se si verifica ciò, l'allele mutato è definito responsabile della patologia ed il gene candidato è definito della suscettibilità alla patologia monogenica.
6. Ogni volta che vengono individuate nuove famiglie con membri portatori della stessa patologia monogenica, viene ripetuta questa stessa strategia partendo dal punto 4. Dalle analisi fatte sul DNA di nuovi malati della stessa patologia si ha la conferma che i membri abbiano mutato un allele dello stesso gene, se sia presente eterogeneità allelica o eterogeneità genica (per ulteriori dettagli vedere punto 7 della strategia della clonazione dei geni candidati funzionali dal capoverso " Per individuare le mutazioni si determina ... .").

Strategia della clonazione posizionale dei geni

Schema della strategia della clonazione posizionale dei geni:
ricerca ed identificazione dei marcatori genetici associati alla patologia monogenica di interesse → costruzione di una mappa genetica densa di marcatori della regione cromosomica associata alla patologia monogenica di interesse → determinazione della sequenza nucleotidica della regione cromosomica mappata che contiene un unico gene → dimostrazione che un allele del gene mappato è mutato → conferma che un allele dello stesso gene è mutato ed associato alla stessa patologia in altre famiglie di aree geografiche e/o di etnie diverse → ricerca della funzione cellulare e dell'attività molecolare della proteina codificata dal gene individuato.

La strategia della clonazione posizionale permette di clonare un gene della suscettibilità ad una patologia monogenica individuando più marcatori genetici strettamente associati alla patologia.

L'esistenza del gene della suscettibilità alla patologia monogenica è dedotta dalla presenza in una o più famiglie di pazienti portatori della patologia monogenica di interesse. I sintomi della patologia (fenotipo patologico) che suggeriscono la presenza dell'allele mutato di un gene ancora ignoto e che sono utilizzati per ricercare i marcatori genetici del gene stesso, forniscono informazioni molecolari insufficienti per candidare il gene come candidato funzionale. In genere non si riesce ad ipotizzare l'attività molecolare e la funzione cellulare della proteina da esso codificata (es. carcinoma della mammella causato dalla mutazione del gene BRCA1 e malattia del " apparente eccesso di mineralcorticoidi", AME).
1. Devono essere ricercati ed individuati più marcatori genetici della patologia. Un allele di ogni marcatore deve essere associato stretto ai malati ed il suo omologo ai sani della stessa famiglia.
Per individuare i marcatori genetici del gene responsabile di una patologia un procedimento consiste nel vagliare la presenza di più marcatori genetici di ciascun cromosoma umano nei portatori di una patologia monogenica appartenenti ad una o più famiglie. L'indagine rivelerà su quale cromosoma mappa la patologia e sopratutto quali alleli del marcatore sono associati alla patologia. Successivamente la ricerca può essere estesa ad altri marcatori aventi il locus sullo stesso cromosoma ed aventi una frazione di ricombinazione inferiore cioè geneticamente più vicini all'allele responsabile della patologia ed essere associati stretti ad esso, al fine di circoscrivere la regione dove mappa la patologia. La ricerca è finalizzata ad individuare più marcatori stretti della patologia, marcatori che rispetto alla patologia non abbiano ricombinato mai dopo un analisi del DNA di più di 100 malati, generati da gameti prodotti da meiosi informative (in genere per analizzare il DNA di 100 malati di una stessa patologia genetica occorre analizzare i membri di decine di famiglie). Con i marcatori genetici si costruisce nella regione cromosomica dove mappa la

patologia (regione candidata), una mappa genetica più dettagliata e piccola possibile al fine che essa includa un solo gene.
2. Il gene è clonato vagliando una genoteca genomica. Il primo clone è isolato utilizzando una PCR analitica che amplifica la sequenza locus-specifica di un marcatore della patologia. Quindi si continua con la tecnica del "camminare sul cromosoma" (figura 4-2).
Negli anni '90 sono state costruite le prime banche elettroniche per conservare le sequenze di DNA genomico e sono stati sviluppati programmi di gestione per poter analizzare le sequenze conservate perché dato l'alto numero di sequenze e di basi che le costituivano era impossibile procedere con i soli mezzi umani. Pertanto la sequenza ottenuta camminando sul cromosoma, è confrontata elettronicamente, mediante programmi di gestione (es. BLAST), con le sequenze conservate nelle banche dati al fine di verificare se in esse si trovi la sequenza di un gene noto o di un nuovo gene. Se il gene è noto, si passa al punto 3. Se la sequenza del gene non è ancora nelle banche dati occorre individuare la sequenza del gene nella sequenza della regione cromosomica sequenziata. Individuare un gene in una sequenza di DNA vuol dire stabilire la sua struttura, cioè l'ordine lineare delle sue sequenze (regione promotrice, esoni, introni e 3'UTR). Ed è necessario dimostrare che il gene è trascritto in un mRNA che codifica una proteina. La migliore prova dell'esistenza di un gene è la dimostrazione che nelle cellule è presente la proteina da esso codificata.
Inizialmente si confronta elettronicamente la sequenza della regione cromosomica di interesse con le sequenze EST umane e di mammiferi nell'apposito archivio elettronico (vedere mappa delle sequenze EST figura 3-14). In esso può essere conservata una sequenza EST, anche di ignota funzione, che sia espressa dal gene di interesse oppure una sequenza EST animale di un gene ortologo. L'ibridazione elettronica con una EST rivela una sequenza genomica codificante e quindi il gene. Utilizzando la sequenza delle EST si possono costruire i primer specifici per la PCR e clonare il cDNA per determinare la struttura del gene mediante l'allineamento della sequenza del cDNA con quella del gene (figura 1-20).
Nel caso in cui l'analisi elettronica delle EST non dia risultati, si usa il procedimento molecolare, simile a quello elettronico. Si vagliano le genoteche di cDNA di differenti tessuti umani ed in particolare dell'encefalo che è l'organo che esprime il maggior numero di geni, circa 20.000 contro i circa 5.000 espressi dagli altri organi. Il vaglio delle genoteche di cDNA è fatto utilizzando la tecnologia della PCR analitica. I suoi primer sono sintetizzati con la sequenza complementare a tratti della sequenza cromosomica di interesse. Si possono avere risultati negativi (nessun clone è individuato) se casualmente si sono costruiti dei primer specifici per la sequenza appartenente ad un introne del gene mappato. Quei primer non possono individuare nessun clone in una genoteca di cDNA. In questo caso, si sintetizzano primer specifici per altri tratti della sequenza mappata fino ad individuare un clone.
Il vaglio della genoteca di cDNA può essere fatto anche utilizzando come sonda il DNA di frammenti della regione cromosomica candidata.

L'inserto di cDNA, contenuto nel clone positivo al vaglio, indica che nella regione mappata è incluso un gene per il fatto che esprime un mRNA. Determinata la sequenza del cDNA, dal suo confronto con la sequenza del DNA della regione mappata si ottiene la struttura in esoni ed introni del gene di interesse.
Nel caso che i precedenti tentativi di individuare (con le EST e con le genoteche di cDNA) il gene di interesse nella regione cromosomica candidata siano stati negativi, si può individuare il gene confrontando elettronicamente la sua sequenza nucleotidica con quelle consenso, comuni a più geni ma non a tutti, conservate nelle banche dati. Queste sequenze includono: sequenze promotrici (es. TATA e CAAT box), sequenze segnale per lo splicing, sequenze segnale di poliadenilazione (es. AAUAAA). Questo tipo di analisi elettronica fatta con le sequenze consenso, a differenza di quella fatta con una sequenza EST (la cui sequenza è unica nel genoma) forniscono solo un suggerimento che deve essere poi necessariamente confermato con tecnologie molecolari. Usando una PCR quantitativa specifica per la sequenza di un esone del gene di interesse si esegue un analisi (simile al Northern) di vari tessuti e si individua il tessuto in cui si ha la più alta espressione del gene. Dal tessuto si estrae lo mRNA totale, si sintetizza il cDNA totale e con esso si costruisce una genoteca di cDNA. Il cDNA di interesse è clonato, usando la stessa PCR usata per l'analisi PCR-Northern, e poi sequenziato.
Il confronto della sequenza del gene con quella del cDNA permette di definire con sicurezza la struttura in esoni-introni del gene di interesse, che è necessaria per effettuare l'analisi al successivo punto 3. Dalla sequenza del cDNA si può dedurre la sequenza aminoacidica della proteina codificata.
Se le analisi per individuare la sequenza e definire la struttura del gene di interesse nella regione cromosomica mappata indicano la presenza di un solo gene si ha l'indicazione che quel gene possa essere quello della suscettibilità alla patologia monogenica.
3. Si verifica che l'allele associato alla patologia abbia la sequenza nucleotidica di uno dei suoi esoni diversa da quella dell'allele/i presente nei membri sani della stessa famiglia. Per questo si determina la sequenza degli amplificati-PCR degli esoni (o si utilizza la tecnica SSCP, vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
Le caratteristiche genetiche dell'allele associato alla patologia devono essere coerenti con quelle della patologia monogenica studiata (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
La presenza nei portatori della patologia monogenica di un allele avente sequenza diversa da quella dell'allele presente nei membri sani è considerata un risultato importante, ma non una prova definitiva, perché l'allele associato alla patologia potrebbe essere una variante non patologica.
4. La conferma dell'identificazione dell'allele mutato come allele patologico e del gene come gene della suscettibilità alla patologia monogenica, può essere data da mutazioni distruttive come le mutazioni non-senso (stop alla traduzione), le delezioni e le inserzioni di basi che portino all'alterazione del quadro di lettura

del messaggero ed in particolare quando la mutazione è vicina al codice di inizio. Le mutazioni distruttive permettono di assumere che la proteina codificata dal gene di interesse sia inattiva, anche senza averne saggiata l'attività molecolare. Mentre le mutazioni missenso, che sostituiscono un aminoacido con un altro, possono essere mutazioni conservative e non patologiche e pertanto occorrono altre prove genetiche e/o molecolari, per avere la certezza che esse siano responsabili di una patologia monogenica.
La conferma definitiva dell'identificazione dell'allele patologico portatore della mutazione missenso ed in senso generale dell'identificazione del gene della suscettibilità alla patologia è ottenuta ripetendo dal punto 3 questa strategia in portatori della stessa patologia monogenica appartenenti ad altre famiglie ed in particolare a famiglie di aree geografiche e/o di etnie diverse. Per verificare la presenza di mutazioni del gene di interesse nei nuovi pazienti si determina la sequenza degli amplificati con PCR eseguite con gli stessi primer usati per individuare il gene responsabile della suscettibilità alla patologia o si esegue la tecnologia SSCP (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
Deve risultare che un allele dello stesso gene è mutato ed associato alla patologia anche nei malati appartenenti ad altre famiglie. Le mutazioni degli alleli nelle diverse famiglie possono essere diverse, per basi diverse nella stessa posizione o per posizioni diverse, causando la sostituzione di aminoacidi diversi nella proteina codificata. In questo caso la patologia monogenica è caratterizzata da eterogeneità allelica. L'eterogeneità allelica è frequente nelle patologie monogeniche.

Se risulta che lo stesso gene, strettamente associato alla patologia monogenica, ha alleli mutati in numerosi pazienti appartenenti a numerose famiglie di aree geografiche e/o di etnie diverse, gli alleli sono definiti responsabili della patologia ed il gene candidato è definito della suscettibilità alla patologia, anche se non ci sono prove indicanti che l'allele mutato codifichi per una proteina alterata nell'attività molecolare e/o nella sua concentrazione cellulare.

Alterazioni della concentrazione della proteina codificata possono risultare da mutazioni del promotore o da mutazioni nel mRNA che modifichino la stabilità della proteina. La presenza di un allele mutato dello stesso gene associato alla stessa patologia monogenica in popolazioni appartenenti ad aree geografiche diverse che non hanno avuto contatti (matrimoni misti) tra loro per molti secoli indica che le mutazioni del gene sono avvenute indipendentemente e prova che le mutazioni causano la stessa patologia monogenica perché interessano lo stesso gene.
Etnie diverse talvolta hanno alleli tipici (etnici), che non sono o sono molto meno presenti in altre etnie, perché gli individui di una data etnia preferiscono sposarsi tra di loro. Quando un allele etnico si trova associato ad una patologia monogenica, pur non essendo responsabile della stessa, ciò può trarre in

inganno il ricercatore che valuti l'associazione di quell'allele etnico con la patologia solo in famiglie di quella etnia. Tuttavia l'analisi dell'allele dello stesso gene in etnie diverse, esclude la possibilità di aver preso un allele etnico al posto di quello patologico.
Individuare un allele mutato di un dato gene associato ad una patologia monogenica in una data etnia e poi riscontrare che un allele mutato dello stesso gene è associato alla stessa patologia monogenica in pazienti di altre etnie (popolazioni che hanno altri alleli etnici) è una conferma importante dell'identificazione di un gene della suscettibilità ad una patologia monogenica.
Può risultare che in una popolazione l'allele etnico sia anche quello responsabile della patologia e che la stessa patologia genetica sia presente con alta incidenza anche nella popolazione di un'altra etnia. In genere ciò dipende dal fatto che nelle due popolazioni gli alleli patologici appartengono a geni diversi.
In questo modo viene dimostrato che la patologia è dotata di eterogeneità genica (eterogeneità di locus) e che è provocata da cause diverse (proteine diverse mutate) e quindi è potenzialmente sensibile a farmaci diversi (vedere patologia MODY2 e appendice E).
5. Ulteriori conferme dell'identificazione del gene della suscettibilità alla patologia monogenica si ottengono individuando la funzione cellulare e l'attività molecolare della proteina da esso codificata. Inoltre la conoscenza della funzione cellulare e dell'attività molecolare della proteina è necessaria per identificare i meccanismi della patogenesi molecolare e per stabilire possibili cure della patologia.

Ricerca della funzione cellulare del gene di interesse:
a) informazioni indirette sulla funzione cellulare della proteina codificata dal gene di interesse si possono ottenere ricercando elettronicamente nelle banche-dati sequenze consenso promotrici di nota funzione (es. sensibili a molecole segnale come gli ormoni steroidei) simili a quelle del promotore del gene di interesse. Altre informazioni si possono ottenere con analisi elettroniche nelle banche dati delle sequenze consenso aminoacidiche di nota funzione (domini funzionali, sequenze segnale, siti di modificazioni post-traduzionali, ecc.) simili a quella della proteina dedotta dal gene di interesse. Le informazioni essendo indirette (non ottenute da analisi fatte sul gene e sulla proteina di interesse) devono essere poi confermate con analisi molecolari (capitolo 1).
b) la ricerca della funzione cellulare della proteina codificata dal gene di interesse può essere effettuata con le tecnologie per la manipolazione dei geni: transfezione in cellule coltivate o in animali modello (es. M. musculus), gene reporter, mutazione sito-specifica, distruzione del gene in animali modello, geni antisenso, mRNA-i, microarray (capitolo 2). La ricerca di geni responsabili di patologie umane in animali modello non sempre fornisce indicazioni utili, perché i geni omologhi negli animali possono avere livelli e tessuti di espressione diversi e la loro mutazione può provocare alterazioni diverse da quelle osservate nell'uomo (capitolo 1).
Ricerca dell'attività molecolare della proteina codificata dal gene di interesse:

Per ricercare l'attività molecolare della proteina occorre sintetizzare la proteina mediante transfezione in cellule di procarioti o eucarioti, purificarla e inventare (cosa che può essere molto difficile) un metodo per saggiare la sua attività molecolare. Le tecnologie descritte sopra ai punti a) e b) danno solo indicazioni indirette sulla possibile attività molecolare della proteina e quindi sul suo dosaggio, pertanto questa ricerca in genere richiede molti tentativi e anni di lavoro di più ricercatori (vedere dopo la clonazione del BRCA1, gene della suscettibilità al carcinoma della mammella). Questa grossa difficoltà a trovare l'attività molecolare della proteina del gene della suscettibilità ad una patologia monogenica clonato con la strategia posizionale, cioè di un gene del quale si conosce solo la sequenza, spiega perché la verifica di alleli mutati dello stesso gene in portatori di una stessa patologia appartenenti a numerose famiglie viene accettata come prova di identificazione del gene della suscettibilità alla patologia.

Strategia della clonazione posizionale dei geni nell'anno 2005

Schema della strategia della clonazione posizionale dei geni nell'anno 2005:
costruzione di una mappa genetica densa di marcatori della regione cromosomica associata alla patologia monogenica di interesse utilizzando marcatori genetici scelti consultando la mappa genetica umana → definizione elettronica della sequenza nucleotidica della regione cromosomica mappata che include un unico gene → dimostrazione che un allele del gene è mutato → dimostrazione che un allele dello stesso gene è mutato ed associato alla stessa patologia in membri di altre famiglie di aree geografiche e/o etnie diverse → ricerca della funzione cellulare e dell'attività molecolare della proteina codificata.

Alcune fasi della strategia della clonazione posizionale dei geni sono rese più semplici dalla possibilità di analizzare elettronicamente dal proprio computer le sequenze dei geni, dei cDNA e delle proteine, le mappe di associazione e fisiche ed altri dati sul genoma umano che sono conservati nelle banche dati (vedere la strategia della clonazione funzionale dei geni 2005).

1. La costruzione di una mappa genetica di ricombinazione ad alta densità di marcatori per ciascuno dei cromosomi umani ha semplificato e reso meno laboriosa la definizione del locus di una patologia monogenica. Per la praticità di analisi (PCR analitica) e per l'alto polimorfismo si preferisce utilizzare almeno inizialmente la mappa genetica dei microsatelliti. Questi marcatori genetici identificano numerosissimi loci per cui definiscono regioni cromosomiche relativamente brevi, mediamente 30kb. Nella banca dati della mappa genetica sono reperibili le sequenze locus specifiche dei microsatelliti marcatori per poterli individuare con la tecnica della PCR analitica.
Quindi inizialmente è sufficiente ricercare con la tecnica della PCR analitica, nei membri malati di una famiglia, la presenza di più marcatori (suggeriti dalla mappa genetica elettronica) con loci distribuiti su tutta la lunghezza di ciascuno

dei 23 cromosomi umani. L'allele di uno dei marcatori dei cromosomi che risulterà presente nel DNA dei malati e assente nei sani della stessa famiglia indicherà il cromosoma sul quale ha il locus l'allele patologico del gene della suscettibilità alla patologia monogenica. Individuato il cromosoma, utilizzando i dati della mappa genetica elettronica, si analizzano marcatori genetici vicini a quello già analizzato al fine di costruire nella regione cromosomica dove mappa la patologia una mappa genetica densa di marcatori associati stretti alla patologia di interesse (vedere punto 1 della strategia della clonazione posizionale dei geni).

2. Costruita la mappa genetica ad alta densità di marcatori strettamente associati alla patologia monogenica e di quelli che allontanandosi da questi hanno una frequenza di ricombinazione maggiore rispetto alla patologia si arriva ad avere una buona definizione della regione cromosomica includente il gene di interesse. A questo punto la sequenza della regione cromosomica può essere ricercata ed individuata esplorando elettronicamente la sequenza del DNA genomico (mappa fisica) dove sono mappati fisicamente anche i marcatori genetici. In relazione alla dimensione fisica della regione mappata e di quella del gene di interesse, la sequenza individuata può includere un unico gene.

3. Si estrae il DNA dei malati e si verifica che l'allele del solo gene presente nella regione mappata sia mutato rispetto all'allele omologo presente nei membri sani della stessa famiglia. Si amplificano gli esoni con la PCR e se ne determina la sequenza (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).

Le caratteristiche genetiche dell'allele associato alla patologia monogenica devono essere coerenti con quelle della patologia studiata (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).

La presenza di un allele diverso del gene di interesse nei portatori della patologia monogenica è considerata una prova importante, ma non definitiva, perché l'allele associato alla patologia potrebbe portare una mutazione conservativa non patologica.

4. La conferma dell'identificazione dell'allele mutato come allele patologico e del relativo gene come gene della suscettibilità alla patologia monogenica, può essere data da mutazioni distruttive che permettano di assumere che la proteina codificata dall'allele sia inattiva, anche senza averne saggiata l'attività molecolare. Mentre le mutazioni missenso possono essere mutazioni conservative e non patologiche, pertanto occorrono altre prove genetiche e molecolari per avere la certezza che siano responsabili di una patologia monogenica.

La conferma definitiva dell'identificazione dell'allele patologico portatore della mutazione missenso ed in senso generale dell'identificazione del gene della suscettibilità alla patologia è ottenuta ripetendo dal punto 3 questa strategia in altre famiglie con membri portatori della stessa patologia monogenica ed in particolare di famiglie appartenenti ad aree geografiche e/o di etnie diverse (per altri dettagli vedere punto 4 della strategia della clonazione posizionale).

Aver individuato il gene responsabile della suscettibilità alla patologia semplifica la ricerca dell'allele patologico in nuove famiglie. Per verificare la presenza di mutazioni del gene di interesse nei nuovi pazienti è sufficiente determinare la sequenza degli amplificati con PCR eseguite con gli stessi primer usati per individuare il gene responsabile della suscettibilità alla patologia (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
5. Ulteriori conferme dell'identificazione del gene della suscettibilità alla patologia possono venire dall'analisi molecolare e dalla funzione cellulare della proteina codificata dal gene della suscettibilità alla patologia.
Se il gene individuato è già stato studiato e nella banche dati si possono trovare le informazioni sulle sue caratteristiche funzionali, e così anche per la proteina codificata, si potrà dosare l'attività molecolare delle proteine mutate sintetizzate *in vitro* e verificare se la loro attività molecolare sia alterata (scarsa o eccessiva) ed avere quindi un'ulteriore conferma che il gene mutato è responsabile della patologia.
Quando la funzione del gene individuato non è nota, la ricerca della funzione cellulare e l'attività molecolare della proteina da esso codificata può essere molto lunga. La conoscenza della funzione cellulare e dell'attività molecolare della proteina è anche necessaria per poter identificare i meccanismi della patogenesi molecolare e per stabilire possibili cure della patologia (per altri dettagli vedere punto 5 della strategia della clonazione posizionale).

Strategia della clonazione dei geni candidati posizionali

Schema della strategia della clonazione dei geni candidati posizionali:
ricerca ed identificazione di marcatori genetici associati alla patologia monogenica di interesse→ mappatura della regione cromosomica associata alla patologia → determinazione della sequenza della regione cromosomica mappata → identificazione della sequenza di più geni candidati nella regione cromosomica associata alla patologia → dimostrazione che un solo gene ha un allele mutato nella regione cromosomica mappata → conferma che un allele mutato dello stesso gene è associato alla stessa patologia in membri di altre famiglie di aree geografiche e/o di etnie diverse → ricerca della funzione cellulare e dell'attività molecolare della proteina codificata dal gene individuato.

La strategia della clonazione dei geni candidati posizionali permette di clonare un gene della suscettibilità ad una patologia monogenica che mappa in una regione cromosomica includente più geni.

1. L'esistenza del gene della suscettibilità alla patologia monogenica è dedotta dalla presenza in una o più famiglie di pazienti portatori della patologia. Tuttavia i sintomi clinici della patologia non forniscono sufficienti informazioni da permettere di ricercare la proteina alterata responsabile della patologia

monogenica, pertanto non può essere applicata la strategia della clonazione funzionale.
Il procedimento per mappare il gene di interesse è identico a quello descritto al punto 1 della clonazione posizionale, ma la regione cromosomica che include il gene di interesse è geneticamente e fisicamente più lunga ed include più geni. I marcatori della regione cromosomica candidata ad includere l'allele responsabile della patologia all'inizio della ricerca possono avere distanze genetiche anche di 20cM. La mappatura della regione può essere migliorata trovando altri marcatori fino a circoscrivere il locus della patologia in una regione inferiore ad 1cM che tuttavia può includere più geni. La mappatura genetica di una patologia può dare indicazioni molto approssimative delle reali distanze fisiche tra i marcatori posti agli estremi della regione cromosomica mappata. Si è osservato che la mappatura può individuare una regione cromosomica dove i marcatori ai suoi estremi hanno una frequenza di ricombinazione relativamente bassa come 1,5%, equivalente ad una distanza genetica di 1,5cM e teoricamente uguale ad 1,5Mb. La reale distanza fisica della stessa regione cromosomica era di 5Mb. Pertanto una regione 1cM può includere più geni candidati posizionali ed alcuni di essi possono costituire un aplotipo all'interno del quale la ricombinazione può essere ferma da decine di secoli.
2. Per ricercare il gene candidato posizionale si determina la sequenza della regione cromosomica dove mappa la patologia, mediante la tecnica del "camminare sul cromosoma", utilizzando inizialmente come sonda la parte locus-specifica di uno dei marcatori genetici della regione cromosomica.
L'identificazione della sequenza dei geni presenti nella sequenza della regione cromosomica candidata è la stessa di quella descritta al punto 2 della strategia della clonazione posizionale dei geni, con la differenza che l'operazione è resa più laboriosa perché i geni da identificare possono essere anche più di 10 in relazione alla dimensione della regione cromosomica mappata ed alle dimensioni dei geni presenti in essa.
3. Si verifica che l'allele di ogni gene candidato presente nella regione cromosomica associata alla patologia abbia la sequenza nucleotidica diversa da quella dell'/degli allele/i presente/i nei membri sani della stessa famiglia. Per far questo si determina la sequenza degli amplificati-PCR degli esoni o si utilizza la tecnica SSCP (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
Dopo questa ricerca, deve risultare che l'allele di uno dei geni candidati ha la sequenza diversa da quella degli alleli dello stesso gene presenti nei membri sani della stessa famiglia e deve avere caratteristiche genetiche coerenti con quelle della patologia monogenica di interesse (per altri dettagli vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
La presenza di un allele diverso nei portatori della patologia monogenica è considerata una prova importante, ma non definitiva, perché l'allele associato alla patologia potrebbe aver subito una mutazione conservativa ed essere una variante non patologica.

4. La conferma dell'identificazione dell'allele mutato come allele patologico e del relativo gene come gene della suscettibilità alla patologia monogenica, può essere data da mutazioni distruttive che permettono di assumere che la proteina codificata dall'allele sia inattiva, anche senza averne saggiata l'attività molecolare. Mentre per dimostrare che le mutazioni missenso, che possono essere mutazioni conservative non patologiche, sono responsabili della patologia monogenica occorre operare altre analisi genetiche e molecolari.
La conferma definitiva dell'identificazione dell'allele patologico portatore della mutazione missenso ed in senso generale dell'identificazione del gene della suscettibilità alla patologia è ottenuta dimostrando che un allele mutato dello stesso gene è associato ai malati appartenenti a famiglie di aree geografiche e/o di etnie diverse (per altri dettagli vedere punto 4 della strategia della clonazione posizionale).
Aver individuato il gene responsabile della suscettibilità alla patologia semplifica la ricerca dell'allele patologico in nuove famiglie. Per verificare la presenza di mutazioni del gene di interesse nei nuovi pazienti è sufficiente determinare la sequenza degli amplificati con PCR eseguite con gli stessi primer usati per individuare il gene responsabile della suscettibilità alla patologia (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
5. Ulteriori conferme dell'identificazione del gene della suscettibilità alla patologia possono venire dall'analisi dell'attività molecolare e della funzione cellulare della proteina codificata dal gene della suscettibilità alla patologia.
Se il gene individuato è già stato studiato e nella banche dati si possono avere le informazioni sulle sue caratteristiche funzionali e così anche sulla proteina codificata, si potrà dosare l'attività molecolare delle proteine mutate sintetizzate *in vitro* e verificare se la loro attività sia alterata (scarsa o eccessiva) e quindi avere la conferma che il gene mutato è responsabile della patologia.
Quando il gene individuato non è noto, la ricerca della funzione cellulare e dell'attività molecolare della proteina da esso codificata può essere molto lunga. La conoscenza della funzione cellulare e dell'attività molecolare della proteina è anche necessaria per identificare i meccanismi della patogenesi molecolare e per stabilire possibili cure della patologia (per altri dettagli vedere punto 5 della strategia della clonazione posizionale).

Strategia della clonazione dei geni candidati posizionali nell'anno 2005

Schema della strategia della clonazione dei geni candidati posizionali 2005:
mappatura della regione cromosomica associata alla patologia monogenica di interesse utilizzando marcatori genetici scelti consultando la mappa genetica umana → definizione elettronica della sequenza della regione cromosomica mappata che include più geni → dimostrazione che un solo gene, che mappa nella regione cromosomica candidata, ha un allele mutato → conferma che un allele dello stesso gene è mutato in portatori della stessa patologia appartenenti

ad altre famiglie di aree geografiche e/o di etnie diverse → ricerca della funzione cellulare e dell'attività molecolare della proteina codificata dal gene individuato.

Alcune fasi della strategia della clonazione dei geni candidati posizionali sono rese più semplici dalla possibilità di analizzare elettronicamente dal proprio computer le sequenze dei geni, dei cDNA e delle proteine, le mappe di associazione e fisiche ed altri dati sul genoma umano che sono conservati nelle banche dati (vedere la strategia della clonazione funzionale dei geni 2005).

1. La costruzione della mappa genetica di ricombinazione ad alta densità di marcatori microsatelliti per ciascuno dei cromosomi umani ha molto semplificato la definizione del locus di una patologia monogenica. Consultando la mappa genetica umana, inizialmente si scelgono più microsatelliti marcatori per ciascuno dei cromosomi umani. L'analisi di questi marcatori nel DNA nucleare dei portatori della patologia appartenenti ad una o più famiglie permetterà di valutare l'associazione dei microsatelliti associati alla patologia e quindi di individuare il cromosoma su cui mappa la patologia. Successivamente, se necessario, si analizzano altri marcatori genetici aventi con la patologia la minore frequenza di ricombinazione (per altri dettagli vedere punto 1 della clonazione posizionale 2005).

La ricerca dei marcatori deve essere fatta necessariamente con analisi genetico-molecolari e non elettronicamente perché della patologia conosciamo solo i sintomi e non ancora la sequenza del gene o della proteina responsabili della stessa patologia. Inoltre anche se il gene della suscettibilità alla patologia e le mutazioni patologiche fossero già note e volessimo analizzare la stessa patologia in una nuova famiglia, sarebbe necessario fare le stesse analisi genetico-molecolari per valutare se nella nuova famiglia il gene responsabile della patologia è lo stesso o un altro (eterogeneità genetica della patologia), oppure se il gene è lo stesso ma è diversa la mutazione (eterogeneità allelica della patologia).

2. Individuati i marcatori genetici che individuano la regione cromosomica più piccola associata alla patologia, si ricerca elettronicamente la sequenza della regione cromosomica nella mappa fisica della sequenza del DNA genomico guidati dagli stessi marcatori genetici, che sono tutti mappati anche fisicamente.

Individuata la sequenza della regione cromosomica mappata, nella stessa pagina elettronica sono mostrate, allineate con essa, le strutture in esoni ed introni dei geni contenuti nella regione mappata, tutti geni candidati posizionali. Ogni gene ha una sigla di riconoscimento attivabile elettronicamente per avere gli altri dati del gene (sequenza, cDNA, proteina codificata, ecc.)

3. Si verifica se l'allele di ogni gene candidato presente nella regione cromosomica associata alla patologia monogenica, abbia la sequenza nucleotidica diversa da quella dell'allele dello stesso gene presente nei membri sani della stessa famiglia. La verifica è fatta determinando la sequenza degli amplificati-PCR degli esoni (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).

Dopo questa ricerca deve risultare che l'allele di uno dei geni presenti nella regione cromosomica associata alla patologia, ha la sequenza diversa da quella degli alleli dello stesso gene presenti nei membri sani della stessa famiglia.
L'allele deve avere caratteristiche genetiche coerenti con quelle della patologia monogenica di interesse (per altri dettagli vedere punto 6 della strategia della clonazione del gene candidato funzionale).
La presenza di un allele diverso nei portatori della patologia monogenica è considerata una prova importante ma non definitiva, perché l'allele associato alla patologia potrebbe aver subito una mutazione conservativa ed essere una variante non patologica.
4. La conferma dell'identificazione dell'allele mutato come allele patologico e del relativo gene come gene della suscettibilità alla patologia monogenica, può essere data da mutazioni distruttive che permettono di assumere che la proteina codificata dall'allele sia inattiva, anche senza averne saggiata l'attività molecolare. Mentre per provare che mutazioni missenso, potendo esse essere mutazioni conservative non patologiche, siano responsabili della patologia monogenica occorre effettuare altre analisi genetiche e/o molecolari.
La conferma definitiva dell'identificazione dell'allele patologico portatore della mutazione missenso ed in senso generale dell'identificazione del gene della suscettibilità alla patologia è ottenuta ripetendo dal punto 3 questa strategia in altre famiglie con membri portatori della stessa patologia monogenica ed in particolare di famiglie appartenenti ad aree geografiche e/o di etnie diverse. La ricerca nelle nuove famiglie è fatta necessariamente con tecnologie genetico-molecolari per la possibile presenza nelle nuove famiglie di eterogeneità allelica o genica. Per verificare la presenza di mutazioni del gene di interesse nei nuovi pazienti si determina la sequenza degli amplificati con PCR eseguite con gli stessi primer usati per individuare il gene responsabile della suscettibilità alla patologia (vedere punto 5 della strategia della clonazione dei geni candidati funzionali).
5. Ulteriori conferme dell'identificazione del gene della suscettibilità alla patologia possono venire dall'analisi molecolare e funzione cellulare della proteina codificata dal gene della suscettibilità alla patologia.
Se il gene individuato è già stato studiato e nella banche dati si possono avere le informazioni sulle sue caratteristiche funzionali e così anche sulla proteina codificata, si potrà dosare l'attività molecolare delle proteine mutate sintetizzate *in vitro* e verificare se la loro attività molecolare sia alterata (scarsa o eccessiva) e quindi avere la conferma che il gene mutato è responsabile della patologia.
Quando il gene individuato non è noto, la ricerca della funzione cellulare e l'attività molecolare della proteina da esso codificata può essere molto lunga. La conoscenza della funzione cellulare e dell'attività molecolare della proteina è anche necessaria per identificare i meccanismi della patogenesi molecolare e per stabilire possibili cure della patologia (per altri dettagli vedere punto 5 della strategia della clonazione posizionale).

Identificazione del gene della suscettibilità alla patologia MODY2

Un esempio dell'applicazione della strategia della clonazione dei geni candidati funzionali.

L'eleganza e semplicità del processo logico seguito dagli scienziati per identificare un gene della suscettibilità alla patologia MODY2 suggeriscono di utilizzare questa ricerca come esempio della strategia della clonazione dei geni candidati funzionali. Questa ricerca è stata diretta da Daniel Cohen, il costruttore della prima mappa fisica dei contig.
La patologia MODY (Maturity-Onset Diabetes of Young) è una delle forme di NIDDM (Non-Insuline Dependent Diabetes Mellitus), diabete mellito non dipendente da una carente sintesi di insulina. MODY è una patologia autosomica dominante che si manifesta in giovane età mentre la forma più comune di NIDDM ha un insorgenza in età avanzata. I pazienti MODY hanno un ridotto rilascio di insulina quando a loro venga somministrato glucosio e ciò aveva suggerito la presenza di un difetto nelle cellule-beta del pancreas dove l'insulina è sintetizzata e rilasciata nel sangue.
I ricercatori proposero che due geni potevano essere candidati funzionali della suscettibilità alla patologia MODY:
1- il gene che codifica la proteina glucosio transferasi che trasporta il glucosio dal sangue nelle cellule-beta del pancreas (GLUT2);
2- il gene che codifica l'enzima glucocinasi (GLK) che catalizza la fosforilazione del glucosio a glucosio-6-fosfato (G6P) utilizzando ATP.
L'ipotesi si basava su alcuni dati ben documentati:
a) la transferasi GLUT2 e l'enzima GLK sono ambedue localizzati nelle cellule-beta del pancreas e l'enzima GLK anche nelle cellule epatiche umane.
b) i due enzimi pancreatici funzionano in serie (GLUT2 --> GLK) ed hanno la funzione di regolare la secrezione dell'insulina in risposta a variazioni del glucosio ematico facendo variare la concentrazione del G6P nelle cellule-beta del pancreas e, tramite il G6P, il rilascio dell'ormone. Un incremento della glicemia (concentrazione di glucosio nel sangue) porta ad un incremento del G6P delle cellule-beta del pancreas e questo ad un incremento del rilascio di insulina. L'insulina ha la funzione di ridurre la glicemia favorendo l'assunzione del glucosio da parte di tutti i tessuti, fegato incluso. Il meccanismo è reversibile come in tutti i sistemi di regolazione: quando la glicemia si abbassa nelle cellule-beta del pancreas, minori quantità di glucosio sono trasformate in G6P. Il G6P cellulare, che è continuamente metabolizzato, rapidamente diminuisce in concentrazione e ciò causa la riduzione del rilascio di insulina nel sangue. Anche la concentrazione dell'insulina nel sangue diminuisce perché continuamente degradata.
c) le caratteristiche molecolari, catalitiche e di regolazione degli enzimi GLK, pancreatico ed epatico, sono identiche. I due enzimi sono codificati dallo stesso gene e per alternative splicing differiscono per pochi aminoacidi all'aminoterminale, mentre la loro funzione fisiologica è molto diversa.

La glucocinasi epatica ha la funzione di sottrarre il glucosio in eccesso quando la glicemia raggiunge livelli alti e di convertirlo in glicogeno e trigliceridi. Gli organi extra-epatici (escluse le cellule adipose) sono privi dell'enzima GLK e convertono il glucosio in G6P per azione dell'enzima esocinasi. Questo enzima ha un'alta affinità per il glucosio (basso valore di Km-glucosio), per cui è saturato dal glucosio (quindi sempre attivo) anche alle concentrazioni ematiche fisiologiche minime; tuttavia l'enzima GLK è inibito da G6P che aumenta in concentrazione quando il metabolismo energetico delle cellule è a livelli ottimali (alti livelli fisiologici di ATP). Gli enzimi GLK (epatico e pancreatico) hanno un valore di Km per il glucosio molto più alto di quello dell'esocinasi per cui non sono mai saturati dal glucosio anche ad alte concentrazioni ematiche fisiologiche; inoltre essi non sono inibiti da G6P pertanto la loro attività catalitica varia solo con il variare dei livelli della glicemia. Risulta così che il fegato è l'unico organo capace di metabolizzare grandi quantità di glucosio in eccesso alle necessità energetiche e sintetiche delle cellule epatiche e di quelle di tutti gli altri organi perché ha le vie metaboliche per convertire il glucosio in molecole di riserva energetica per l'intero organismo. Queste molecole sono il glicogeno epatico e gli acidi grassi e trigliceridi che saranno esportati e conservati prevalentemente nelle cellule adipose.
Nel caso che il gene-GLK fosse responsabile del MODY, il ridotto o mancato funzionamento dell'enzima GLK epatico dovrebbe contribuire all'iperglicemia e quindi al MODY perché il glucosio in eccesso non potrebbe essere sottratto al sangue.
d) I pazienti MODY mostrano intolleranza al glucosio, cioè dopo somministrazione orale di glucosio hanno livelli di glicemia maggiori di 7,8mM, valore stabilito dall'Organizzazione Mondiale della Sanità come indice di intolleranza al glucosio. Questa osservazione ha favorito l'ipotesi che uno dei due geni candidati, se mutato, potesse essere responsabile della patologia MODY, perché l'intolleranza al glucosio dipende dalla mancata attivazione del meccanismo molecolare di rilascio nel sangue dell'insulina da parte delle cellule-beta del pancreas che causa l'iperglicemia tipica del diabete mellito.
I ricercatori conclusero che una mutazione del gene-GLUT2 o del gene-GLK potesse causare un insufficiente rilascio di insulina e quindi una ridotta assunzione di glucosio da parte dei tessuti e del fegato con la conseguente iperglicemia osservata nei pazienti MODY.
Se uno dei due geni fosse stato responsabile della suscettibilità alla patologia MODY, esso doveva risultare associato alla patologia ed essendo la patologia MODY dominante, un allele variante di quel gene doveva essere presente solamente nei pazienti MODY e mai nei membri sani della stessa famiglia.
Al fine di verificare quanto sopra ipotizzato, i ricercatori clonarono il gene-GLK. Il cDNA-GLK era stato già clonato e sequenziato da altri ricercatori mediante clonazione funzionale. I ricercatori del gruppo di Cohen sintetizzarono due primer complementari alla sequenza del cDNA-GLK e mediante PCR analitica vagliarono una genoteca genomica umana, camminarono sul cromosoma ed identificarono il gene-GLK. Quindi stabilirono la sua struttura in esoni ed introni

mediante allineamento delle due sequenze (del cDNA e del gene, figura 1-20). Il gene-GLK è costituito da 12 esoni ed è lungo più di 20kb (figura 4-4b). Poi ricercarono i marcatori del gene-GLK. Un microsatellite (CA) chiamato GLK1 (3 alleli) che marcava il gene-GLK sul Chr 7p13 fu trovato nella letteratura scientifica, era già stato individuato da ricercatori giapponesi. Un altro microsatellite marcatore della glucocinasi, chiamato GLK2 (quattro alleli), fu individuato da un ricercatore del gruppo di Cohen perché il microsatellite GLK1 non era informativo in tutte le famiglie che avevano analizzato.
Il microsatellite GLK1 era stato individuato dai ricercatori giapponesi camminando sul cromosoma (nelle due direzioni), utilizzando come prima sonda un frammento del cDNA-GLK. Il microsatellite GLK1 fu individuato a circa 10kb dal 3' dal gene-GLK. Utilizzando il metodo della analisi genetica delle cellule somatiche ibride ed il marcatore GLK1 come sonda, il gene-GLK fu mappato sul cromosoma umano 7; successivamente altri ricercatori mapparono il gene-GLK sulla banda p13 dello stesso cromosoma 7. La distanza fisica di 10kb corrisponde teoricamente alla lunghezza genetica di 0,01cM (frequenza di ricombinazione 0,01%). Ciò permise di assumere che il marcatore GLK1 fosse strettamente associato al gene-GLK e quindi poteva essere utilizzato per l'analisi dell'associazione del gene con la patologia. Nel caso che il microsatellite non fosse stato associato stretto al gene le successive analisi di associazione tra marcatore e gene avrebbero escluso o confermato la loro associazione. L’associazione fu confermata (vedere dopo) con una frequenza di ricombinazione maggiore (0,1%) di quella teorica, calcolata sopra. Comunque l'associazione risultava essere sufficientemente stretta per ricercare il gene di interesse, teoricamente distante 1Mb dal locus del marcatore.
Anche il microsatellite $(CA)_n$ che marcava il gene-GLUT2 sul Chr 3q26.1-3 fu trovato nella letteratura (era stato individuato da altri ricercatori e pubblicato su una rivista scientifica). Il microsatellite era localizzato all'interno dell'introne-4a del gene-GLUT2, quindi era sicuramente associato al gene.
I ricercatori utilizzando i marcatori sopra indicati, verificarono l'associazione genetica del gene-GLUT2 e del gene-GLK con la patologia MODY in 16 famiglie francesi aventi almeno due membri portatori della patologia MODY, per un totale di 125 individui malati.
Il locus della patologia MODY non era ancora noto, ma ciò non limitava la ricerca, perché la possibile associazione tra il locus del marcatore ed il locus del gene responsabile della patologia (ancora ignoto) poteva essere verificata osservando che lo stesso allele del marcatore passava da un genitore malato ai figli malati e cosi avanti nelle generazioni successive delle famiglie analizzate. Essendo la patologia MODY dominante doveva risultare che un dato allele di uno dei due geni era presente solo nei malati ed assente nei sani. Nella stessa famiglia, gli alberi genealogici dei portatori di quell'allele dovevano risultare identici a quelli dei portatori della patologia. Fu estratto il DNA dalle cellule del sangue dei 125 malati e per controllo di 2 sani, e mediante PCR analitica furono analizzati il marcatore del gene-GLUT2 ed i marcatori GLK1 e GLK2 del gene-GLK. Fu calcolata la frazione di ricombinazione, il numero di ricombinanti fu

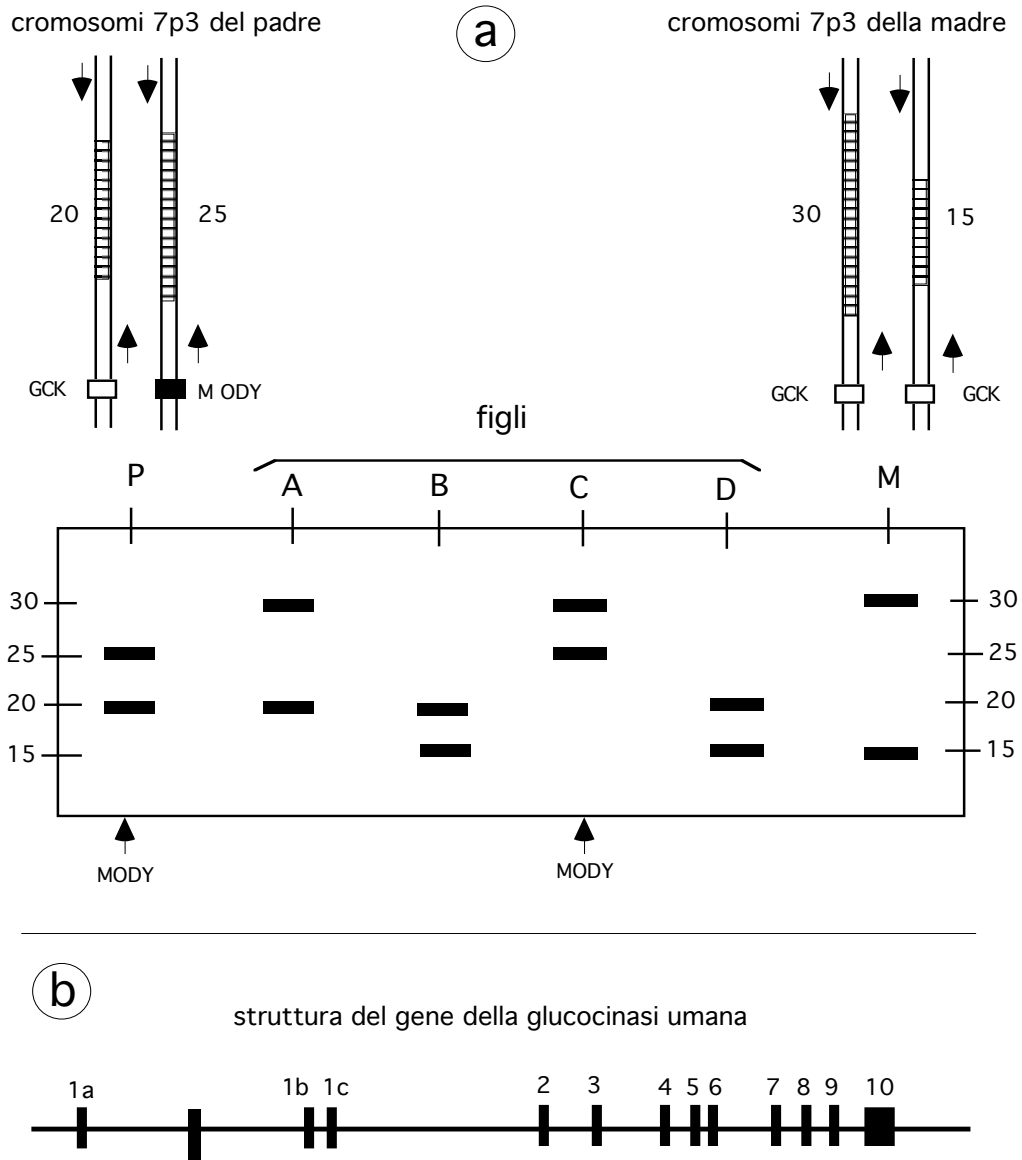



Figura 4-4
a. Associazione del gene dell'enzima glucocinasi (GLK) alla patologia MODY. La trasmissione dell'allele con 25 ripetizioni in tandem del marcatore del gene-GLK dal padre P malato al figlio A malato, dimostra che il gene-GLK è associato alla patologia e che l'allele con 25 ripetizioni in tandem del marcatore è diagnostico della patologia MODY in quella famiglia. Gli altri microsatelliti con 15, 20 e 30 ripetizioni in tandem marcano la presenza dell'enzima glucocinasi normale. Eventuali ricombinazioni tra il gene-GLK e il gene della patologia sarebbero mostrate dalla presenza del microsatellite con 25 ripetizioni in tandem in membri sani della stessa famiglia ed il marcatore non sarebbe più informativo. Le coppie di frecce parallele ai cromosomi indicano i primer per la PCR che associandosi a regioni, uniche ed identiche nel genoma di tutti gli individui della popolazione umana, identificano il locus del marcatore associato stretto al gene-GLK.
b) struttura del gene umano dell'enzima glucocinasi. Il trascritto primario del gene è soggetto a splicing alternativo. Lo mRNA pancreatico è costituito dal trascritto degli esoni 1a e 2-10 e lo mRNA epatico principalmente dal trascritto degli esoni 1b e 2-10 e scarsamente dagli esoni 1b, 1c e 2-10.

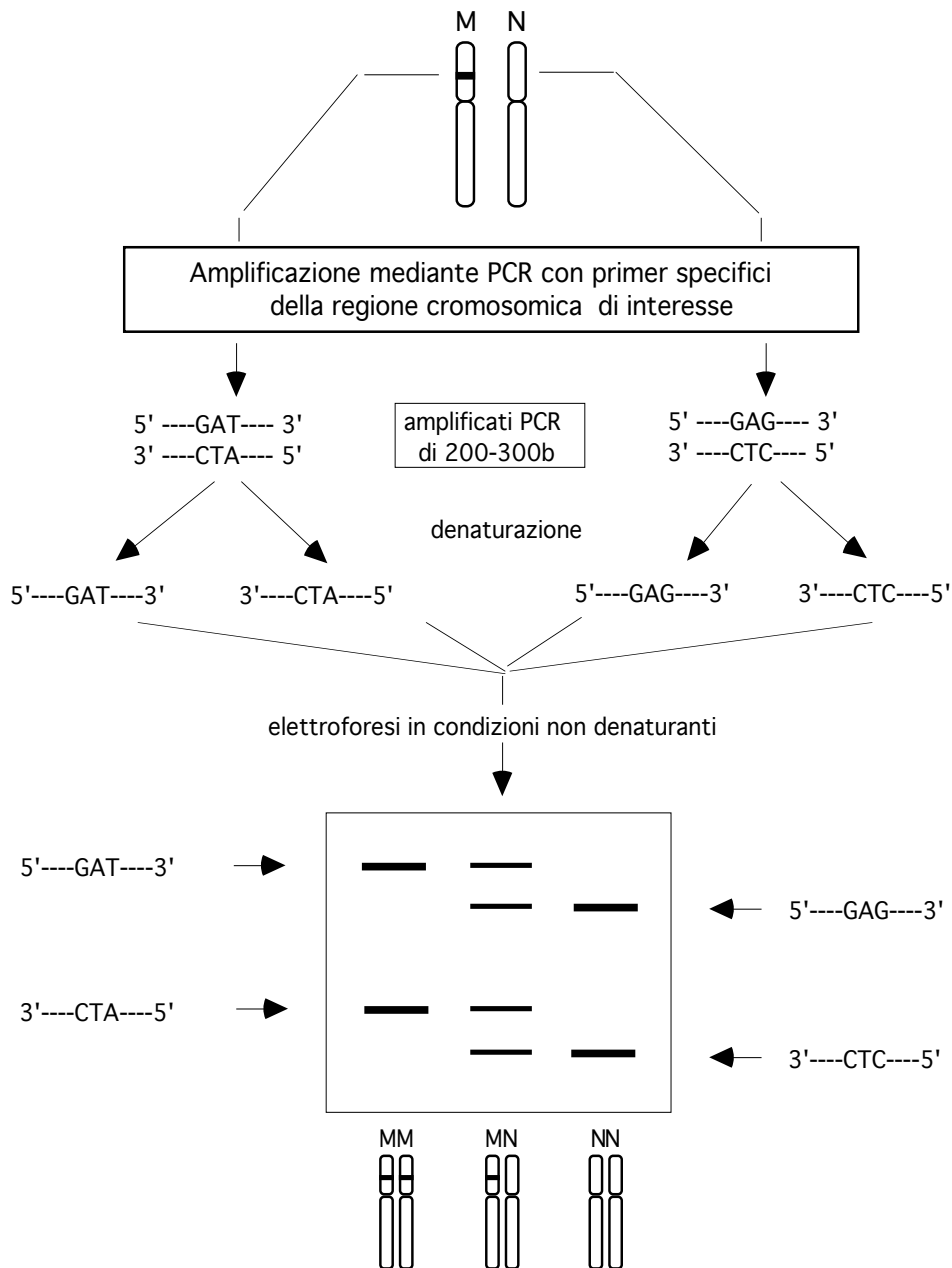



Figura 4-5. Descrizione schematica della tecnica del polimorfismo della conformazione dei singoli filamenti (SSCP, Single-Strand Conformation Polymorphism). Il DNAds della regione cromosomica di interesse (200-300b) viene amplificato mediante PCR, gli amplificati sono denaturati e sottoposti subito ad elettroforesi su gel di poliacrilamide in condizioni non denaturanti. I singoli filamenti di DNA (DNAss) all'inizio dell'elettroforesi assumono spontaneamente e poi mantengono conformazioni diverse in relazione alla loro sequenza per cui i filamenti senso e antisenso di uno stesso amplificato di DNA migrano in posizioni diverse ed i filamenti mutati senso ed antisenso migrano in posizioni diverse, che sono anche diverse da quelle dei filamenti normali omologhi. Le bande sono rivelate mediante autoradiografia (se è utilizzato un NTP radioattivo per la sintesi del DNA) oppure mediante colorazione con i sali di argento. MM = Mutato omozigote, NN = Normale omozigote MN = eterozigote. In alto della figura sono indicati solo i cromosomi di un individuo eterozigote MN. Per altri dettagli vedere testo.

(ridisegnato e modificato da Recombinant DNA, Watson J.D., Gilman M., Witkowski J. and Zoller M. (1992) Recombinant DNA, 2nd ed., Scientific American Books, Freeman, USA).

diviso per il numero di tutti gli individui malati analizzati (generati da gameti provenienti da meiosi informative) e la validità della frazione di ricombinazione controllata calcolando il Lod score. L'analisi genetica rivelò che nessun allele del gene-GLUT2 era associato alla patologia: i dati statistici di associazione non erano significativi, il Lod score era negativo nelle 12 famiglie esaminate.
I marcatori del gene-GLK avevano con la patologia MODY una frazione di associazione uguale a 0,01 che corrisponde ad una frequenza di ricombinazione del 1% convalidata da un Lod score = 11,6 nelle 16 famiglie analizzate. Questa analisi dimostrò che i due marcatori genetici del gene-GLK erano anche associati stretti alla patologia con una distanza genetica tra marcatore e la patologia (cioè all'allele mutato del gene patologico ancora ignoto) di 1cM.
I risultati dell'analisi dell'associazione tra i marcatori del gene-GLK e la patologia dimostrarono che gene-GLK e la patologia MODY mappavano sulla stessa banda p13 del Chr7, pertanto il gene dell'enzima glucocinasi, candidato funzionale, era ora anche candidato posizionale. Tuttavia non si poteva escludere che un altro gene fisicamente vicino al gene-GLK potesse essere responsabile di MODY. Comunque il risultato ottenuto era un passo avanti molto importante, se non ci fosse stata l'associazione genetica tra il gene della glucocinasi e la patologia MODY, la responsabilità del gene-GLK nella patologia sarebbe stata esclusa come era stato fatto per il gene-GLUT2.
L'analisi dell'associazione genetica fu fatta (e si continua a fare) con i marcatori e non con gli alleli (normale e mutato) del gene di interesse perché l'uso dei marcatori rende più semplice e più veloce la ricerca dell'associazione del gene con la patologia. Ricercare il polimorfismo del gene candidato funzionale prima di aver stabilito l'associazione di un suo allele alla patologia può essere un lavoro inutile come dimostrato dalla esclusione dalla candidatura del gene-GLUT2. Inoltre individuare il polimorfismo di un microsatellite è molto più semplice (una analisi PCR per ogni gene) che analizzare il polimorfismo di un gene, il che richiede di analizzare tutti gli esoni del gene perché le mutazioni possono essere su esoni diversi di malati appartenenti a famiglie diverse, come accade anche per MODY (tabella 4-1). Il gene-GLK ha 12 esoni (figura 4-4b), occorre quindi moltiplicare almeno per 12 le analisi per ogni individuo analizzato. La ricerca degli alleli richiede ancora più analisi se il gene è mutato sul promotore.
PCR per ogni gene) che analizzare il polimorfismo di un gene che richiede di analizzare tutti gli esoni del gene perché le mutazioni possono essere su esoni diversi di malati appartenenti a famiglie diverse, come accade anche per MODY (tabella 4-1). Il gene-GLK ha 12 esoni (figura 4-4b), occorre quindi moltiplicare almeno per 12 le analisi per ogni individuo analizzato. La ricerca degli alleli richiede ancora più analisi se il gene è mutato sul promotore.
I dati funzionali della glucocinasi e l'associazione genetica del gene-GLK alla patologia MODY favorivano l'ipotesi che il gene della glucocinasi fosse responsabile della suscettibilità alla patologia MODY ed indussero a verificare se l'allele del gene-GLK, associato ai malati di MODY, avesse una sequenza mutata e quindi diversa da quella dell'allele della glucocinasi dei membri sani della

stessa famiglia. Per dimostrare la presenza della possibile mutazione i ricercatori utilizzarono la tecnica detta del "polimorfismo della conformazione dei singoli filamenti" o tecnica SSCP (Single-Strand Conformation Polymorphism, figura 4-5).
Per applicare la tecnologia SSCP, i ricercatori amplificarono con PCR segmenti di DNA di circa 300b coprendo l'intera sequenza di ciascuno dei 12 esoni (le dimensioni dei quali sono comprese tra 90 e 234 basi) del gene della glucocinasi di 125 pazienti MODY appartenenti a famiglie diverse e per controllo fecero le stesse analisi sugli esoni di due membri sani delle stesse famiglie per un totale di oltre 1500 analisi SSCP.
La tecnica SSCP è una particolare elettroforesi che permette di evidenziare la presenza di mutazioni in frammenti di coppie di alleli omologhi.
Il DNAds, amplificato con la PCR dalla coppia di alleli omologhi, viene denaturato ed i filamenti di DNAss tendono ad avvolgersi su loro stessi per assumere conformazioni sequenza-specifiche, stabilizzate soprattutto da legami ad H tra le basi. I due filamenti complementari di DNAss amplificati da uno stesso frammento di DNAds, avendo sequenza diversa, assumono conformazioni diverse. Dopo la denaturazione, i filamenti di DNAss sono caricati subito su un gel di elettroforesi avente condizioni non-denaturanti (assenza di molecole denaturanti e di alte temperature). Alcuni filamenti di DNAds si riassoceranno a formare di nuovo il filamento di DNAds originale, altri filamenti di DNAss assumeranno le loro conformazioni sequenza-specifiche che manterranno anche trascinati dalla corrente elettrica. La conformazione tridimensionale rallenta la corsa dei filamenti di DNAds in maniera diversa in relazione alla conformazione assunta dai filamenti di DNAss che è sequenza specifica. Mentre i filamenti di DNAds che si sono riformati migreranno più velocemente ed in relazione alla loro lunghezza (figura 1-6).
La sensibilità della tecnologia è sufficientemente alta da separare le strutture autoassemblate dei filamenti di DNAss di 200-300b amplificati da coppie di alleli che hanno sequenze che differiscono anche per una sola base. Questo è molto importante perché, utilizzando la tecnica della PCR, il DNA estratto da un individuo diploide come l'uomo, porta ad amplificare segmenti di DNA cromosomico di ambedue gli alleli. In condizioni di eterozigosi si formano 4 filamenti diversi che appaiono come quattro bande nell'elettroforetogramma della tecnologia SSCP. Negli individui omozigoti sani si vedranno solo due bande corrispondenti ai due filamenti senso ed antisenso di DNAss che corrispondono a regioni di esoni dei due alleli, paterno e materno, aventi la stessa sequenza. Negli individui portatori della mutazione MODY, che in genere è in eterozigosi perché la patologia è dominante, si vedranno quattro bande: due corrispondenti ai filamenti di DNAss (senso ed antisenso) dell'allele normale e due corrispondenti ai filamenti di DNAss (senso ed antisenso) dell'allele portatore della mutazione puntiforme (figure 4-4a e 4-5).
Analizzando con la tecnica SSCP il DNA diploide prelevato dai pazienti MODY appartenenti a più famiglie, fu osservato che tutti i pazienti avevano un allele variante del gene-GLK. In relazione alla famiglia analizzata, la variazione di

sequenza dell'allele variante era localizzata su un dato esone o nelle basi (AG e GT) dei punti di taglio dello splicing (figura 1-20). In totale gli esoni varianti trovati nelle diverse famiglie furono 9 (tabella 1-3). Ogni paziente MODY analizzato aveva un esone variante, e l'esone variante era lo stesso nei malati della stessa famiglia, mentre nessuno degli esoni varianti era presente nei membri sani della stessa famiglia.
Delle varianti di sequenza degli esoni 5, 6, 7, 8 di alcuni pazienti MODY furono individuate le basi mutate mediante analisi della sequenza degli amplificati PCR, che data la diploidia, includevano due differenti forme alleliche di DNA.
I ricercatori ottennero un'ulteriore conferma costruendo una genoteca di ogni amplificato-PCR dei vari esoni, isolarono più cloni (alcuni cloni avevano il DNA dell'allele mutato, altri di quello normale) e determinando la sequenza dei vari cloni.
Il confronto di queste sequenze con quelle omologhe dei soggetti sani permise di individuare le basi diverse nei malati di MODY (tabella 4-1).
A questo punto si era fatto un altro importante passo avanti verso l'identificazione del gene responsabile di MODY. Si era dimostrato che alleli del gene-GLK, presenti solo nei pazienti MODY, avevano una sequenza diversa da quella degli alleli presenti nei sani; tuttavia queste varianti potevano essere mutazioni missenso conservative, cioè non patologiche.
Occorreva quindi dimostrare che gli alleli varianti codificavano glucocinasi cataliticamente scarse o non attive.
Una chiara indicazione che l'enzima GLK mutato fosse responsabile del MODY, era venuta dall'identificazione di una mutazione missenso amber (Glu-->stop) nell'esone 7 del gene-GLK di pazienti MODY di una famiglia.
L'incompleta sintesi della proteina enzimatica indicava che almeno in quel paziente una assente attività catalitica glucocinasica era strettamente associata alla patologia MODY e quindi doveva essere responsabile di essa. Si poteva sempre obiettare che la mutazione era nella parte terminale del gene e che la proteina, sebbene incompleta, poteva essere ancora attiva.
Per le altre forme varianti del gene-GLK i ricercatori costruirono dei plasmidi di espressione operando nel modo seguente:
- un plasmide includeva come inserto il cDNA dell'enzima epatico che differisce per splicing-alternativo da quello pancreatico per pochi aminoacidi al N-terminale (figura 4-4b).
- un plasmide includeva come inserto il cDNA dell'enzima pancreatico.
- varie copie dell'inserto dell'enzima pancreatico furono modificate mediante mutazione sito specifica (vedere capitolo 2, figura 2-3) in modo da avere vari plasmidi con inserti di cDNA-pancreatico con le stesse mutazioni riscontrate nei pazienti.
I plasmidi con gli inserti dei cDNA normali e mutati furono transfettati in Escherichia coli e gli enzimi espressi furono purificati fino alla omogeneità.
Il cDNA con la mutazione amber non produceva nessuna proteina. Si assume che la proteina, essendo sintetizzata più corta, sia subito degradata perché meno resistente alle proteasi batteriche per la sua incompleta struttura

Glucocinasi	Vmax (unità/mg di enzima)	Km (mM) glucosio	ATP
isoenzimi normali:			
fegato	98	6,8	0,23
pancreas cellule-beta	100	8	0,15
isoenzimi mutati:			
1 Gly-175-->Arg	51	39	0,10
2 Val-182-->Met	49	70	0,20
3 Val-203-->Ala	0,5	100	0,20
4 Thr-228-->Met	0,4	10	0,20
5 Glu-256-->Lys	0,25	2,4	0,20
6 Gly-261-->Arg	0,46	2,5	0,20
7 Glu-279-->Amber	inattivo	-	-
8 Glu-279-->Gln	55	41	0,20
9 Gly-299-->Arg	0,32	3,1	0,15
10 Glu-300-->Lys	33	25	0,14
11 Glu-300-->Gln	100	20	0,19
12 Leu-309-->Pro	0,98	2,2	0,13

Tabella 4-1. Caratteristiche cinetiche dell'enzima glucocinasi normale e mutato.
Il numero a tre cifre indica la posizione nella sequenza della proteina dell'aminoacido sostituito dalla mutazione.
Gli isoenzimi che portano i numeri:
- 4 e 5 hanno l'aminoacido mutato localizzato nella cavità di legame del glucosio
- 6-11 hanno l'aminoacido mutato vicino alla cavità di legame del glucosio
- 12 ha l'aminoacido mutato localizzato lontano dal sito catalitico dell'enzima
- 6, 9, 12 hanno l'aminoacido mutato che provoca una distorsione nella struttura dell'enzima
- le mutazioni: n° 1 e 2 sono nell'esone 5; n° 3 nell'esone 6, n° 4-8 nell'esone 7; n° 9-12 nell'esone 8.
(da Gidh-Jain M. et al. Proc Natl Acad Sci U S A, 90,1932-6, 1993, parzialmente modificato).

terziaria. Tutti gli enzimi portanti mutazioni missenso (eccetto la mutazione-GLU300) avevano valori di Vmax molto inferiori rispetto agli enzimi normali. La riduzione della Vmax fu attribuita ad una distorsione della struttura terziaria della proteina e/o ad incrementi della Km per il glucosio causati da alterazioni del sito di legame del glucosio.
La mutazione-Glu300 porta, in contrasto con altre mutazioni, ad un alto valore di Km per il glucosio ma non della Vmax. Ciò è spiegato considerando che la Vmax è misurata in condizioni di saturazione di glucosio, cioè a concentrazioni di glucosio alte non raggiungibili anche in condizioni di alta glicemia fisiologica.
Pertanto in condizioni fisiologiche di glicemia, la velocità di catalisi della GLK mutata-GLU300 risulta essere i nsufficiente a garantire nelle cellule-beta del pancreas la sintesi di G6P per una efficiente regolazione del rilascio di insulina al crescere della glicemia.
Questi risultati dimostrarono che l'enzima glucocinasi, codificato dall'allele mutato del gene-GLK presente solo nei malati, era cataliticamente poco o non attivo e pertanto era responsabile della patologia MODY.
I risultati ottenuti erano in accordo con altri dati ed osservazioni sperimentali.

Lo stesso gene-GLK codifica le due glucocinasi (epatica e pancreatica), le due proteine differiscono per splicing alternativo dell'esone all'amino-terminale, esone non interessato da mutazioni patologiche (4-4b e tabella 4-1). L'enzima GLK pancreatico svolge una funzione di regolazione ed anche una
piccola variazione della sua attività catalitica ha conseguenze drammatiche nel rilascio dell'insulina. L'attività catalitica della glucocinasi nelle cellule-beta è critica nello stabilire la soglia alla quale deve essere rilasciata l'insulina in risposta ad incrementi di glucosio ematico. E' stato calcolato che una modesta riduzione del 15% dell'attività glucocinasica sposta la soglia di rilascio di insulina da 5mM a 6mM. I livelli normali della glicemia sono tra 4,44 e 6,1mM. Pertanto mutazioni di un singolo allele (eterozigosi) che causino la perdita parziale di attività glucocinasica di almeno il 15% sono sufficienti ad instaurare la patologia MODY ed a provocare l'iperglicemia non fisiologica. La minore secrezione di insulina provoca una minore attività di importazione e di metabolismo del glucosio in tutti i tessuti, fegato incluso. Inoltre la scarsa o nulla attività catalitica della GLK epatica che nei malati MODY porta la stessa mutazione nell'enzima GLK pancreatico, aggrava lo stato clinico dei pazienti. Questa stessa ricerca ha anche confermato che gli enzimi facenti parte di meccanismi molecolari di regolazione, se mutati, causano patologie dominanti (vedere appendice D).
L'insieme dei dati sperimentali ottenuti utilizzando la strategia della clonazione dei geni candidati funzionali, hanno dimostrato che il gene della glucocinasi è quello responsabile della suscettibilità alla patologia MODY.
Con la clonazione del gene e le prove sull'attività molecolare della proteina codificata si sono chiarite: la causa della patologia (alterazione genetica della proteina) e la patogenesi molecolare, cioè come l'alterazione genetica dell'enzima glucocinasi provochi l'alterazione della regolazione del rilascio di insulina e quindi l'iperglicemia, uno dei sintomi del diabete MODY.
Per la sua responsabilità nella patologia MODY, la glucocinasi è considerata un importante bersaglio farmacologico. Si ipotizza che riuscendo a sintetizzare farmaci capaci di attivare specificamente l'enzima glucocinasi dell'allele normale dei malati di MODY, si dovrebbe riuscire ad annullare gli effetti della patologia restaurando il corretto rilascio dell'insulina.
Studi successivi hanno mostrato che la stessa patologia MODY è causata anche da altri geni e quella causata dal gene che codifica l'enzima glucocinasi è stata chiamata MODY2.
Gli altri geni della suscettibilità alla patologia MODY sono: il gene del fattore 4-alfa nucleare del fegato (Chr20, MODY1); il gene del fattore di trascrizione-1 epatico (Chr 12q24, MODY3); il gene del fattore-1 del promotore dell'insulina (Chr13q12.1, MODY4); il gene del fattore di trascrizione-2 epatico (Chr17 MODY5); il gene NEUROD1 (Chr2q32, MODY6).

Identificazione del gene BRCA1 responsabile della suscettibilità all'insorgenza precoce del carcinoma della mammella.

Un esempio dell'applicazione della strategia della clonazione dei geni candidati posizionali.

Nel 1988 era stato dimostrato che nel 4% delle donne l'insorgenza del carcinoma della mammella era attribuibile a fattori ereditari e colpiva donne con età inferiore ai 45 anni (insorgenza precoce del carcinoma della mammella). Nel 1990, fu individuata l'associazione dell'insorgenza precoce del carcinoma della mammella con il cromosoma 17 banda q21 (Chr17q21) e furono fatti i primi tentativi per individuare in quella regione cromosomica il gene responsabile del carcinoma. Successivamente, dopo un lungo lavoro di ricerca di scienziati di più laboratori, è stato dimostrato che responsabile della suscettibilità all'insorgenza precoce del carcinoma della mammella era un gene che fu chiamato BRCA1 (**BR**east **CA**ncer gene **1**).
I ricercatori, all'inizio della pubblicazione scientifica che nel 1990 riportava l'identificazione della regione cromosomica 17q21, spiegarono la logica della loro strategia per clonare il gene, poi chiamata della clonazione dei geni candidati posizionali ed utilizzata per clonare molti altri geni della suscettibilità a patologie. La traduzione di quanto scrissero è:
"***il locus dei geni umani patologici può essere individuato mediante l'analisi dell'associazione di famiglie in cui la patologia ha una alta incidenza. L'analisi di associazione può rivelare la localizzazione cromosomica dei geni di interesse per mezzo dell'identificazione dei marcatori genetici polimorfici di noto locus che nelle famiglie sono coereditati con la patologia***".
I ricercatori analizzarono 23 famiglie abitanti in Porto Rico, Canada, Colombia, Gran Bretagna ed in 40 stati USA. Furono analizzate 329 donne delle quali 146 portatrici di tumore della mammella. Utilizzando complessivamente 183 marcatori genetici polimorfici fu analizzata l'associazione dei marcatori di ogni singolo cromosoma con la patologia. L'analisi individuò un marcatore VNTR (D17S74), un minisatellite altamente polimorfico che mappava sulla banda 21 del braccio lungo del cromosoma 17 (Chr17q21). Il minisatellite fu analizzato digerendo il DNA nucleare delle pazienti con l'enzima di restrizione HinfI che tagliava la regione di DNA nucleare, includente il minisatellite D17S74, in frammenti di lunghezza diversa in relazione alle ripetizioni in tandem di ogni allele del marcatore D17S74. Il marcatore D17S74 ha più di 30 alleli e le loro lunghezze variano da 1 a 5kb. I frammenti includenti il marcatore D17S74 erano poi individuati mediante un'analisi Southern utilizzando una sonda avente sigla CMM86, specifica per locus del minisatellite D17S74.
Nell'analisi delle 23 famiglie fu osservato che un particolare allele del marcatore era associato alle portatrici del carcinoma aventi meno di 45 anni di età (insorgenza precoce), ma non alle portatrici del tumore che si era formato più tardivamente. I ricercatori riscontrarono nella regione mappata la presenza di più geni candidati posizionali ed alcuni di questi con caratteristiche di possibili

candidati funzionali. Questi geni codificavano un fattore di crescita cellulare, un enzima che catalizzava la sintesi dell'estradiolo (ormone promotore della proliferazione cellulare), un recettore dell'acido retinoico (che ha un ruolo nello sviluppo embrionale). Successivamente fu dimostrato che nessuno di questi geni era il gene della suscettibilità al carcinoma della mammella.
Nel 1994, altri ricercatori utilizzando dei marcatori microsatelliti da loro individuati, costruirono una mappa genetica ad alta risoluzione di circa 40cM intorno al minisatellite D17S74 marcatore della regione Chr 17q21. Cinque di quei microsatelliti furono individuati determinando con sequenziatore automatico la sequenza della regione cromosomica. La sequenza del DNA della regione cromosomica mappata e quelle dei marcatori furono conservate in una banca dati locale. Queste due tecnologie, determinazione della sequenza del DNA con un sequenziatore automatico e conservazione delle sequenze nucleotidiche in archivi elettronici corredati di programmi di gestione, erano all'epoca importanti novità che permettevano un rapido confronto delle sequenze, altrimenti impossibile.
Successivamente gli stessi ricercatori, dopo aver ristretto 4cM la mappa della regione cromosomica 17q21 associata al carcinoma della mammella, costruirono di essa anche la mappa fisica che aveva ai due suoi estremi due microsatelliti da loro individuati. La mappa completa era costituita da un contig formato dagli inserti di DNA genomico umano di 137 cloni YAC con sequenze in parte sovrapponibili ed includeva 112 nuovi marcatori fisici analizzabili con la tecnica PCR (successivamente chiamati STS).
Nella mappa fisica furono individuati 20 geni di cui 10 erano già noti e mappati geneticamente nella regione. Gli stessi ricercatori analizzarono due di questi geni posti in una regione ipotizzata essere quella che includeva il gene ricercato, ma non riuscirono a trovare in essi delle mutazioni (successivamente si dimostrò che nessuno dei due geni era il gene BRCA1).
La competizione in questa ricerca era molto forte (come nelle officine delle moto che partecipano ai gran premi internazionali) e sempre nel 1994, ricercatori di un altro laboratorio costruirono indipendentemente una mappa fisica di 2,3Mb della regione Chr17q21 associata al carcinoma della mammella e compresa tra due marcatori genetici: D17S776 prossimale (più vicino al centromero) e D17S78 distale. Il contig di 2,3Mb includeva inserti di cloni provenienti da più genoteche genomiche umane aventi come vettori: il batteriofago P1 (29 inserti), YAC (11 inserti) e l'inserto di un cosmide. Il contig includeva 51 sequenze STS, 9 nuovi microsatelliti polimorfici ed 8 geni noti. Utilizzando i nuovi marcatori genetici la localizzazione del gene BRCA1 fu ristretta ad una regione di 600kb delimitata dai marcatori microsatelliti D17S1321 prossimale e D17S1325 distale.
Al fine di individuare i geni della regione di 600kb, furono vagliate le genoteche umane di cDNA di mammella, ovaia, linfociti, encefalo fetale ed altri tipi di sequenze espresse utilizzando come sonde frammenti diversi del DNA della regione cromosomica mappata. I cDNA clonati e sequenziati, avendo la loro sequenza identica a quella della sonda di DNA nucleare, indicarono le sequenze

degli esoni dei geni presenti nella regione mappata. Furono clonate 65 sequenze espresse (poi chiamate EST) candidate posizionali, gruppi di esse che appartenevano ad uno stesso gene (figura 3-13) furono utilizzate per costruire THC candidati che includevano interamente il quadro di lettura (orf), cioè l'intera parte del mRNA che codifica la proteina. Lo mRNA, poi identificato per quello trascritto dal gene del BRCA1, fu costruito da tre sequenze espresse, esso appariva completo nella parte codificante avendo al 5' la sequenza consenso di Kozak di inizio della traduzione ed al 3' la sequenza segnale di poliadenilazione (AUAAA) seguita da un tratto di poli-A. Il cDNA ricostruito codificava una proteina ignota di 1863 aminoacidi. L'allineamento della sequenza del cDNA con quella del DNA genomico permise di stabilire che il gene BRCA1 era costituito da 22 esoni codificanti, distribuiti su circa 100kb. Determinazioni successive hanno stabilito che il gene è lungo 81.090b e gli esoni sono 24 inclusi quelli non codificanti.
Tutti i geni codificanti gli mRNA ricostruiti, furono analizzati nelle portatrici del carcinoma della mammella e si osservò che tra questi geni uno solo aveva alleli varianti presenti solo nelle portatrici di BRCA1.
In pazienti appartenenti a famiglie diverse furono trovate mutazioni di tipo diverso: una delezione di 11 coppie di basi, un'inserzione di 1 base, una mutazione nonsenso (sostituzione di una aminoacido con un codone di stop) ed una mutazione missenso.
Nei primi tre tipi di varianti sopra indicati, l'alterazione della sequenza dell'allele era tale da far ritenere che la proteina codificata non potesse avere una normale attività molecolare, cioè erano mutazioni potenzialmente distruttive. La delezione di 11 coppie di basi causava la perdita di 3 aminoacidi, ma soprattutto modificava il quadro di lettura (il numero di basi delete non è divisibile per 3) del mRNA, che dopo la mutazione codificava una proteina con la sequenza completamente diversa dalla proteina normale. Analoga considerazione può essere fatta per la mutazione per inserzione di una base che anch'essa altera il quadro di lettura. La mutazione nonsenso, causando uno stop prematuro alla traduzione del mRNA, portava alla sintesi di una proteina incompleta e presumibilmente inattiva. La mutazione missenso era quella che *di per se* non era indicativa di una alterazione distruttiva della proteina potendo essere una mutazione conservativa che non alterava l'attività molecolare della proteina e quindi non patologica (appendice D).
Le analisi Northern utilizzando il cDNA come sonda mostrarono che l'mRNA-BRCA1 era lungo 7,8kb e che era espresso nella mammella e nell'ovaia ma anche e più abbondantemente nel testicolo e nel timo.
Il confronto della sequenza della proteina BRCA1 con quelle conservate nelle banche dati permise di individuare la sequenza di un dominio "zinc finger" (appendice B) nella sequenza NH_2terminale della proteina BRCA1. Ciò suggeriva che la proteina potesse legarsi al DNA. La capacità della proteina BRCA1 di associarsi al DNAds è stata dimostrata successivamente con tecnologie molecolari.

L'identificazione del gene della suscettibilità precoce al tumore della mammella è stata confermata da analisi condotte su famiglie di molti paesi e di etnie diverse dove un allele mutato del gene BRCA1 è stato trovato associato al carcinoma della mammella in donne di età inferiore ai 45 anni.
Secondo un modello di interpretazione dei dati sperimentali, il gene normale BRCA1 è indicato come gene soppressore dei tumori (antioncogene) responsabile, se mutato, di una carcinogenesi recessiva. Il carcinoma insorge quando la mutazione è omozigote. Tuttavia quando un allele mutato è trasmesso dai genitori alle figlie, sebbene l'allele mutato sia recessivo, la suscettibilità a sviluppare il tumore è trasmessa in maniera dominante. Questo accadrebbe perché le cellule epiteliali della ghiandola mammaria possono spontaneamente divenire omozigoti per l'allele mutato mediante uno dei meccanismi di perdita di eterozigosi (LOH, appendice D). La ghiandola mammaria è formata da milioni di cellule, pertanto esiste un'alta probabilità che si formi il tumore perché è sufficiente che la LOH si realizzi in una singola cellula per instaurare la carcinogenesi. Questa osservazione suggerisce che le cellule della ghiandola mammaria abbiano un corredo proteico che favorisce la LOH del BRCA1 eterozigote e quindi la carcinogenesi a differenza di altri tipi di cellula dello stesso organismo che non sviluppano il tumore. Fanno eccezione le cellule dell'ovaia: si è osservato che in alcune famiglie le portatrici del BRCA1 mutato sviluppano tumori sia nella mammella e che nell'ovaia.
Dopo l'identificazione del gene BRCA1 sono stati condotti molti studi sulla proteina BRCA1. Questi studi hanno mostrato che essa è una fosfo-proteina nucleare coinvolta in alcuni processi cellulari che includono: modificazione della cromatina, associazione al DNA, trascrizione, riparazione del DNA, ricombinazione omologa, poliadenilazione del mRNA, controllo del ciclo cellulare, ubiquitinazione delle proteine. L'ubiquitina è una proteina di 76 aminoacidi, ubiquitaria negli eucarioti, che viene legata covalentemente alle proteine per segnalare ai sistemi proteolitici cellulari quando una proteina deve essere degradata. La proteina BRCA1 normale ha la funzione di mantenere il genoma integro, per questo essa è associata ad altre proteine “soppressori di tumori”, sensori del DNA danneggiato e trasduttori di segnale formando un complesso di proteine chiamato BASC (BRCA1 Associated genome Survegliance Complex).
Solo l'associazione al DNA e l'ubiquitinazione sono state osservate con un dosaggio biochimico che valuta, *in vitro* e con molecole pure, l'attività molecolare della proteina BRCA1. Le altre funzioni sono state osservate in cellule intere, pertanto quelle funzioni possono essere una conseguenza indiretta (tramite altre proteine) dell'attività molecolare della proteina BRCA1. Questi dati sono in accordo con la indicazione che il BRCA1 sia il gene della suscettibilità al tumore della mammella, tuttavia non è ancora ben delineata la catena di eventi molecolari che iniziano per azione della proteina geneticamente modificata e finiscono con l’insorgenza del tumore della mammella.
Mancando una prova molecolare diretta, la migliore indicazione che il gene BRCA1 sia il gene della suscettibilità all'insorgenza precoce del carcinoma della mammella è l'alta frequenza di associazione di un

suo allele mutato con le donne di età inferiore ai 50 anni, portatrici di quel tipo di tumore.

I ricercatori che avevano clonato il gene BRCA1 durante le loro analisi trovarono una donna di 80 anni sana e portatrice del BRCA1 mutato. Statistiche successive hanno mostrato che le donne che hanno un allele del gene BRCA1 mutato hanno un alto rischio (92% dei casi) di sviluppare durante la loro vita il carcinoma della mammella, mentre nelle donne che hanno ambedue gli alleli integri il rischio è ridotto al 10%. L'osservazione che l'8% delle donne, portatrici del gene BRCA1 mutato non sviluppino il carcinoma è spiegata assumendo che l'insorgenza del carcinoma dipenda anche dalla combinazione degli alleli del patrimonio genetico della donna portatrice della mutazione BRCA1. Questi geni esprimono proteine capaci di influenzare la fisiologia delle cellule della ghiandola mammaria, proteine che possono essere sintetizzate nelle stesse cellule della ghiandola mammaria od in altre cellule. A conferma di ciò è l'osservazione che solo in una frazione delle donne portatrici della mutazione BRCA1 insorge sia il carcinoma della mammella che quello dell'ovaia; inoltre, lo stesso gene BRCA1 mutato sebbene espresso in altri tipi di cellula (es. timo) non provoca in quelle cellule l'insorgenza del carcinoma. Altri fattori possono inibire l'insorgenza del carcinoma della mammella: il tipo di alimentazione, l'ambiente, la cultura e lo stile di vita della portatrice della mutazione BRCA1. Analoghe considerazioni possono essere fatte per tutte le patologie che hanno una componente genetica (appendici D, E ed ultime pagine di questo capitolo).

Per ottenere le informazioni sulla proteina è stato necessario il contributo di alcuni anni di lavoro di molti ricercatori attivi in decine di laboratori di paesi diversi. Occorre ricordare che non è semplice ricercare l'attività molecolare della proteina BRCA1, proteina di grosse dimensioni (1863 aminoacidi) con più domini. Tuttavia, la ragione principale del lento progredire della ricerca dell'attività molecolare di una proteina è dato dal fatto che le tecnologie di manipolazione dei geni danno informazioni sulla funzione cellulare delle proteine ma non sulla loro attività molecolare, proprietà che identifica le proteine stesse (capitolo 2). La conoscenza dell'attività molecolare di una proteina permette di individuare i meccanismi molecolari che determinano la sua funzione cellulare. Quando la stessa proteina è mutata, l'analisi della alterazione della sua attività molecolare (eziologia) permette di individuare i meccanismi molecolari responsabili dell'alterazione della funzione cellulare (patogenesi molecolare) e delle loro manifestazioni (sintomi).

Strategie per l'identificazione dei geni responsabili dei fenotipi complessi normali e patologici

I fenotipi complessi normali e patologici dipendono dall'espressione di più geni (fenotipi normali poligenici e patologie poligeniche) non necessariamente posti sullo stesso cromosoma e quindi non necessariamente associati geneticamente.

I fenotipi normali poligenici e le patologie poligeniche sono detti multifattoriali quando la loro manifestazione dipende oltre che dall'espressione di più geni anche da fattori ambientali (fisici, chimici e biologici), alimentari e culturali (vedere appendice E).
I fenotipi complessi normali includono la dimensione dei globuli rossi, il peso e l'altezza corporei, il colore della pelle, la pressione arteriosa, l'intelligenza ed il comportamento (il modo in cui una persona interagisce con altre persone e con l'ambiente). Le patologie complesse includono le neoplasie, il diabete mellito degli adulti, l'ipertensione essenziale, l'epilessia, l'autismo, la schizofrenia ed altre malattie psichiatriche. Le patologie complesse sono definite anche comuni o sociali perché, nella popolazione umana hanno incidenze più alte di quelle delle patologie monogeniche che in genere sono rare.
I fenotipi normali e patologici complessi sono ricorrenti nei membri di una stessa famiglia con una incidenza del 5-10% alla prima generazione (contro il 25-50% dei caratteri e delle patologie monogeniche), pertanto sono definiti non-mendeliani. I singoli geni che partecipano a formare fenotipi normali o patologici ubbidiscono alle leggi di Mendel, ma non le combinazioni dei loro alleli.
Nessuno degli alleli da solo è capace di instaurare il fenotipo complesso normale o la patologia complessa, ma tutti vi concorrono. Inoltre alleli diversi possono avere pesi diversi ed i loro effetti sono additivi nel determinare il fenotipo o la patologia complessa.
Si assume che uno stesso carattere normale poligenico (esempio l'altezza corporea) sia l'espressione di un gruppo di geni, poiché quegli stessi geni possono avere più specie molecolari di alleli. Lo stesso carattere poligenico (esempio l'altezza corporea) può avere variabilità in una popolazione (altezze diverse in individui diversi) in relazione alla combinazione degli alleli presenti in ogni singolo individuo. In maniera analoga, una stessa patologia complessa, in relazione alla combinazione degli alleli del gruppo di geni responsabili della suscettibilità "debole" alla patologia complessa, può avere variabilità di sintomi.
La suscettibilità dei geni responsabili dei fenotipi complessi normali e patologici è detta "debole" perché un solo gene non può essere responsabile di una patologia complessa.
Nei portatori di patologie complesse, l'incidenza di alcune specie molecolari di alleli (alleli della dominanza "debole") del gruppo di geni responsabili della suscettibilità "debole" della patologia, è maggiore che nei soggetti sani.
Gli alleli di uno stesso gene codificano proteine diverse per uno o pochi aminoacidi aventi una normale attività molecolare e una normale funzione cellulare ma diverse proprietà minori o subdole (vedere capitolo 1 e appendice E). Queste proprietà includono valori diversi di attività molecolare, di vita molecolare, di capacità di associare e subire gli effetti di molecole endogene (metaboliti e molecole segnale) e molecole esogene (alimenti, inquinanti, allergeni, farmaci, anestetici). Gli effetti di queste molecole possono esaltare o inibire la normale attività molecolare o la normale funzione cellulare delle proteine. Le proprietà minori delle proteine contribuiscono alla variabilità fenotipica degli individui normali (capitolo 1 e appendice D). Le proprietà minori

sono dette subdole, quando inattese agiscono negativamente sulla fisiologia dell'organismo e si manifestano solo in particolari stati fisiologici (digiuno prolungato, sforzo muscolare, basse pressioni di ossigeno) oppure quando nell'organismo è introdotta una particolare molecola (inquinante, farmaco, ecc.) che si associa ad una proteina e ne altera l'attività molecolare.
Si assume quindi che combinazioni di proteine, aventi particolari proprietà minori o subdole, provochino l'alterazione patologica complessa e che al superamento della soglia patologica possa contribuire anche un fattore ambientale (es. sale nell'ipertensione, fumo del tabacco nelle cardiopatie).
L'analisi degli alleli e quindi delle proteine da essi codificate suggerisce che differenti combinazioni di isoproteine aventi differenti proprietà minori o subdole, codificate da alleli diversi dello stesso gruppo di geni, siano responsabili sia delle funzioni normali dell'organismo umano e delle sue variazioni normali (individui più o meno forti, più o meno intelligenti), sia delle alterazioni delle stesse funzioni in particolari condizioni fisiologiche (es. digiuno prolungato) o di alterazioni permanenti (patologie complesse).

L'identificazione dei geni coinvolti nelle patologie complesse è molto più laboriosa di quella dei geni delle patologie monogeniche per i seguenti motivi:

1 - la dipendenza della patologia da più alleli,
2 - la variabilità della combinazione degli alleli dei geni che causano la stessa patologia,
3 - la difficoltà a distinguere le varianti alleliche coinvolte nella patologia (dette della suscettibilità "debole") dalle altre varianti normali, per l'assenza di reali mutazioni patologiche che possano guidare la ricerca genetica e molecolare.
4 - l'impossibilità di fare analisi di associazione genetica (linkage) tra sintomi e possibili geni, dato che nessun gene ha un allele mutato presente solo o con alta incidenza nei portatori della patologia e assente negli individui sani.
5 - l'identificazione dei geni responsabili di malattie psichiatriche è ancora più difficile perché si conoscono poco o non si conoscono del tutto le alterazioni molecolari che le causano. Inoltre secondo alcuni autori, i criteri diagnostici per classificare le malattie psichiatriche, non essendo fisici o biochimici, non sono biologicamente validi ed utilizzabili per indagini genetiche. Al fine di poter fare analisi genetiche, alcuni ricercatori hanno proposto di associare le malattie psichiatriche a sintomi misurabili quantitativamente come il peso, temperatura corporea e la pressione arteriosa.

La ricerca dei geni coinvolti in patologie complesse risulta più semplice quando l'alterazione interessa la funzione di vie metaboliche o di regolazione delle quali si conoscono le proteine che le costituiscono.
Conoscendo le proteine è facile risalire ai geni che le codificano, analizzare il loro polimorfismo (vedere la clonazione funzionale nel 2005) per poi studiare singolarmente i differenti alleli. Un esempio è dato dalla ricerca diretta ad individuare la componente genetica dell'ipertensione essenziale.

Ipertensione essenziale.

L'ipertensione, un persistente aumento della pressione del sangue arterioso, è definita secondaria quando dipende da cause ben definite delle quali alcune genetiche (es. sindrome di Liddle). L'ipertensione essenziale è primaria e le cause della sua insorgenza non sono interamente conosciute. L'ipertensione essenziale è la forma più comune ed interessa il 90% dei casi. Normalmente si manifesta dopo i 40 anni di età, ha andamento familiare con ereditarietà di tipo poligenico multifattoriale sulla quale incidono fattori ambientali ed alimentari ed in particolare l'eccessiva ingestione di alimenti contenenti ione sodio (Na^+).

Nell'ipertensione essenziale non si osservano rilevanti cambiamenti costanti in alcuno dei parametri (noti) che possono modificare la pressione arteriosa: cioè i livelli di renina, aldosterone, catecolamine o l'alterazione dei barocettori del sistema nervoso simpatico.

Un gruppo di ricercatori francesi e statunitensi, utilizzando la strategia del gene candidato funzionale, ha analizzato 379 famiglie aventi, ognuna di esse, alcuni membri ipertesi. Essi hanno dimostrato che esiste una associazione genetica tra il gene che codifica l'angiotensinogeno e l'ipertensione essenziale e che lo stesso gene è polimorfico con 15 differenti varianti genetiche.

L'angiotensinogeno è una proteina che fa parte della via di regolazione della pressione arteriosa (figura 4-6). Esso è prodotto e rilasciato nel sangue dal fegato e per divenire attivo (angiotensina-2) subisce gli attacchi proteolitici di due enzimi diversi. L'angiotensina-2 ematica è inattivata da un enzima proteolitico specifico pertanto ha un turnover dato da una continua sintesi e degradazione.

Un incremento della concentrazione ematica di angiotensinogeno provoca l'incremento della pressione arteriosa aumentando il riassorbimento nel sangue del Na^+ del filtrato renale. Il Na^+, essendo circondato da molecole d'acqua (acqua di idratazione), entrando nel sangue ne aumenta il volume e l'aumento di volume del sangue causa un incremento della pressione nelle arterie .

I ricercatori osservarono che i pazienti ipertesi avevano valori più alti di concentrazione ematica di angiotensinogeno e che le concentrazioni diverse corrispondevano a specifiche varianti genetiche codificanti angiotensinogeno sintetizzato in quantità maggiore o degradato con minore velocità. Tuttavia quelle stesse varianti erano presenti, anche se in percentuali significativamente inferiori, in membri normotesi della stessa famiglia. Ciò suggerì che combinazioni di alleli di altri geni della via molecolare che regola la pressione arteriosa dovevano contribuire all'insorgere dell'ipertensione essenziale che infatti è una malattia poligenica multifattoriale. Gli stessi ricercatori hanno proposto che le varianti genetiche di angiotensinogeno, che risultano avere una concentrazione ematica più alta, siano uno degli aspetti molecolari della predisposizione ereditaria all'ipertensione per gli individui che le posseggono.

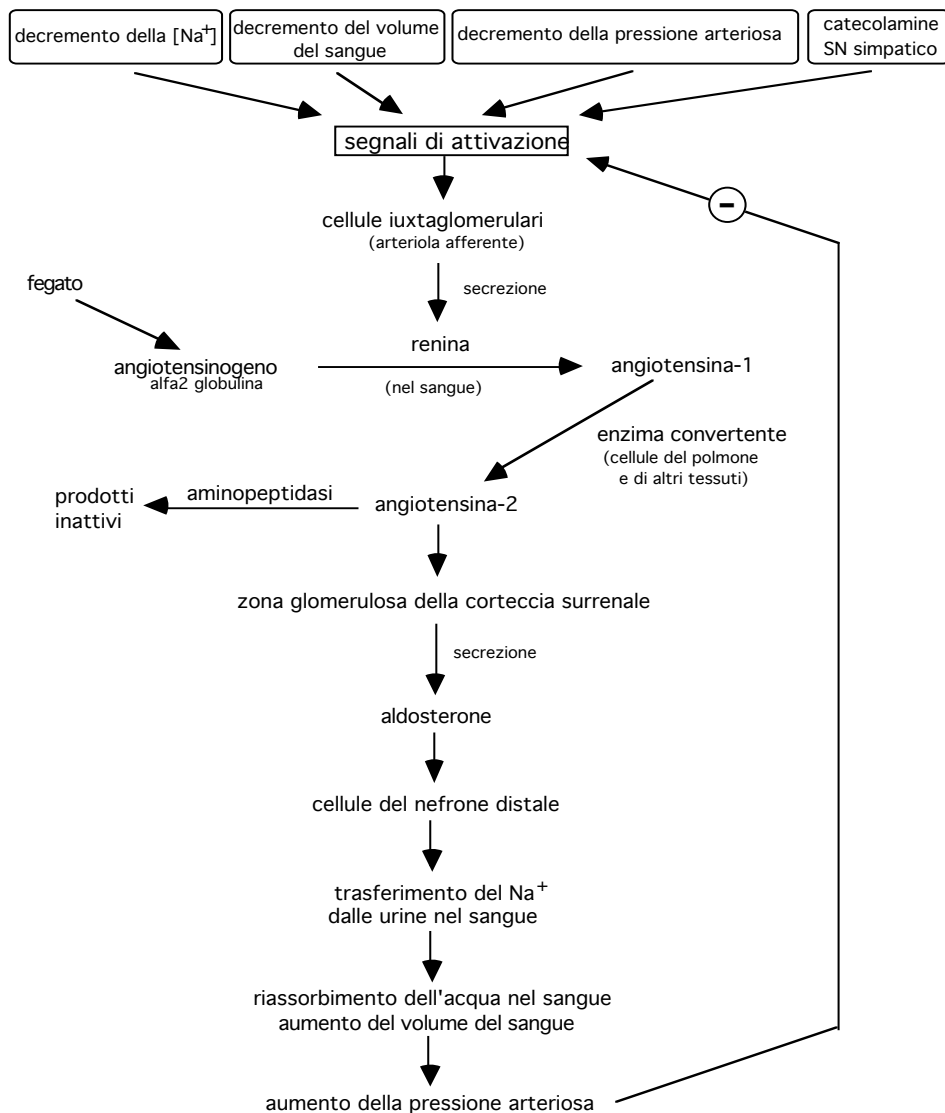



Figura. 4-6. Schema del meccanismo molecolare di regolazione della pressione arteriosa del sistema ghiandola surrenale-rene (ridisegnato e modificato da A.W. Norman and G. Litwack Hormones, 1987 Academic Press, San Diego). Per altri dati vedere testo.

Ricerca dei geni responsabili della suscettibilità a patologie complesse

La ricerca dei geni responsabili delle patologie complesse in alcuni casi è stata fatta analizzando più geni contemporaneamente, ed anche sull'intero genoma umano valutando l'associazione (association) statistica non-mendeliana di alleli appartenenti a geni diversi posti su cromosomi diversi. Questa associazione è definita non-mendeliana perché non è dedotta dalla percentuale di ricombinazione di geni posti sullo stesso cromosoma. In inglese i termini sono diversi, linkage (collegamento) è l'<u>associazione</u> <u>genetica</u> valutata statisticamente di geni posti sullo stesso cromosoma, mentre l'<u>association</u> è

l'associazione anch'essa valutata statisticamente di alleli di geni posti su cromosomi diversi che viene rilevata analizzando le combinazioni patologiche dei loro alleli. Nei malati l'associazione non-mendeliana degli alleli coinvolti nella patologia deve risultare superiore a quella che gli stessi alleli hanno nei membri sani di una stessa famiglia.
Si assume che la suscettibilità a patologie complesse sia determinata principalmente da varianti alleliche geneticamente antiche, come suggerisce la loro grande distribuzione nella popolazione umana, cioè non ristrette ad aree geografiche o a particolari etnie perché le mutazioni sarebbero sorte in progenitori comuni a gran parte della popolazione umana.
Pertanto per alcune ricerche di geni coinvolti in patologie complesse sono state utilizzate le sequenze SNP come marcatori genetici perché più numerose (circa 1 SNP ogni 1000 basi di DNA) e geneticamente più stabili dei microsatelliti. I microsatelliti sono circa 1 ogni 10.000 basi e tendono a mutare più frequentemente dei marcatori SNP. Inoltre, si tende a ricercare marcatori SNP all'interno delle sequenze codificanti (esoniche) del gene, assumendo che tra questi SNP ci siano anche quelli responsabili delle proprietà minori della proteina, e quindi della suscettibilità "debole" alla patologia complessa. In questo modo la base SNP è marcatore dell'allele della dominanza "debole" ed anche responsabile della proprietà minore della proteina codificata. Alcune industrie hanno costruito macchine e preparato un corredo di soluzioni già pronte che permettono di analizzare 10.000 SNP in un giorno.

Strategia della scansione posizionale del genoma per la ricerca dei geni responsabili dei fenotipi complessi normali e patologici

La strategia della scansione posizionale del genoma è utilizzata per identificare gli alleli dei marcatori genetici polimorfici distribuiti su tutto il genoma. La disposizione degli alleli di un numero di marcatori genetici sufficientemente grande da dare la visione dettagliata delle ricombinazioni che avvengono in tutto il genoma è detto "profilo degli alleli del genoma". Questa strategia applicata ai membri appartenenti a più generazioni di una o più famiglie permette di ricercare i geni responsabili dei fenotipi normali complessi ed i geni della suscettibilità alle patologie complesse anche quando di quei geni non siano conosciuti né funzione né locus.
La strategia della scansione posizionale del genoma si basa sull'analisi mediante PCR analitica o micro-sequenziamenti di un alto numero di marcatori sparsi su tutti i cromosomi nucleari di individui portatori di un dato fenotipo complesso (es. un dato comportamento) o di una patologia complessa (es. ipertensione). I profili degli alleli ottenuti con questa prima analisi su almeno due generazioni sono poi confrontati con i profili degli alleli di individui appartenenti alle stesse famiglie rispettivamente aventi un diverso fenotipo complesso (diverso comportamento) o individui sani (es. normotesi), al fine di individuare gli alleli

dei marcatori genetici più frequentemente associati ai portatori di un dato comportamento o di una data patologia.
La scansione posizionale del genoma porta a mappare e quindi ad individuare gli alleli delle regioni cromosomiche aventi una frequenza di associazione non-mendeliana verso un fenotipo normale complesso (es. altezza corporea oltre 1,70 m) o verso una patologia complessa (es. ipertensione arteriosa) statisticamente più alta di quella che gli stessi alleli hanno rispettivamente verso un altro fenotipo normale complesso (es. altezza corporea inferiore a 1,70 m) o verso lo stato di salute (pressione arteriosa normale). Queste regioni cromosomiche sono regioni definite regioni candidate posizionali ed i geni presenti in esse sono detti geni candidati posizionali a determinare quel fenotipo complesso o quella patologia complessa. La maggior frequenza di associazione al fenotipo complesso o alla patologia complessa dell'allele di un gene candidato posizionale suggerisce che quell'allele debba avere un ruolo nel determinare rispettivamente il fenotipo normale complesso o nel causare la patologia complessa. L'analisi degli alleli dei marcatori in almeno due generazioni rivela anche l'origine materna o paterna degli alleli dei geni candidati posizionali, origine che può essere indipendente dal sesso del genitore o esclusiva di un dato sesso del genitore.
Mappate le regioni cromosomiche associate al fenotipo o alla patologia complessa si individuano le loro sequenze ricercandole elettronicamente nelle banche dati contenenti l'intera sequenza del genoma umano. Le stesse banche dati indicheranno i geni contenuti in quelle sequenze.
Se i geni presenti nella regione mappata sono stati già clonati e sono note l'attività molecolare e la funzione cellulare della proteina codificata, si può individuare tra essi il possibile gene candidato funzionale che contribuisce a determinare il fenotipo o la patologia complessa. Se i geni sono ignoti, occorre ricercare l'attività molecolare e la funzione cellulare delle proteine codificate al fine di trovare tra loro un possibile candidato funzionale.
Come indicato sopra la difficoltà ad individuare i geni delle patologie complesse risiede nel fatto che gli alleli che contribuiscono alla patologia non hanno mutazioni distruttive che possano identificarli, essendo essi presenti con frequenza minore anche negli individui sani. Pertanto occorre passare dalla candidatura posizionale (ottenuta mediante scansione posizionale) alla candidatura funzionale. Le differenze di sequenza tra gli alleli di uno stesso gene co-responsabile di una patologia complessa sono determinate da mutazioni missenso conservative (mai distruttive) per cui esse non possono dare *di per se* informazioni sul grado di dominanza "debole" dei singoli alleli. Per avere queste informazioni occorre analizzare il grado di attività delle proprietà minori e subdole e/o di ogni isoproteina codificata da ogni singolo allele del gene di interesse (così come è stato fatto per gli alleli dell'angiotensinogeno).
La scansione posizionale del genoma permette di individuare anche aplotipi associati alle patologie complesse. Gli aplotipi sono ritenuti importanti per l'identificazione di queste patologie perché essi possono includere alleli di geni e di marcatori che non hanno ricombinato per lunghi periodi durante l'evoluzione

umana. Pertanto, quando l'allele di un dato marcatore e un allele di un gene coinvolto in una patologia ed in particolare di una patologia complessa mappano in un dato aplotipo, si ha la certezza che nelle varie generazioni marcatore e patologia siano rimasti associati per molte generazioni. La mappa degli aplotipi umani è stata completata nel dicembre 2005.

La ricerca dei geni responsabili dell'orientamento sessuale degli uomini come esempio della strategia della scansione posizionale

Per questa ricerca sono stati utilizzati 403 microsatelliti aventi loci su tutti i cromosomi umani. La distanza media tra essi è di 7cM (teoricamente 7Mb), distanza che può includere più geni, tuttavia il numero dei loci analizzati è tale da dare un profilo molto dettagliato degli alleli distribuiti in tutto il genoma degli individui di una o più famiglie e quindi delle possibili ricombinazioni.
Utilizzando questi microsatelliti sono stati analizzati 456 individui di 146 famiglie aventi due o tre fratelli omosessuali. Si è analizzata la costituzione allelica dei genitori e dei figli maschi omosessuali ed eterosessuali ed i profili allelici ottenuti sono stati confrontati al fine di vedere se particolari alleli di qualche marcatore erano presenti più frequentemente nel genoma di figli omosessuali rispetto al genoma dei loro fratelli eterosessuali. Poiché i fratelli hanno mediamente in comune il 50% del genoma, gli alleli dei marcatori di regioni cromosomiche includenti geni coinvolti nell'orientamento sessuale devono essere presenti in percentuali superiori al 50%.
Gli alleli dei marcatori genetici più frequentemente presenti nei membri omosessuali marcavano le regioni cromosomiche 7q36 e 8q12, rispettivamente con il 62,50% e 60,10% rispetto al 50% della distribuzione casuale degli alleli di altri geni. Questo è un esempio di associazione non genetica di due geni (i due geni sono posti su cromosomi diversi) tuttavia l'informazione ottenuta con la strategia della scansione posizionale indica che gli alleli dei geni inclusi in quelle regioni cromosomiche possono contribuire all'orientamento sessuale maschile.
L'analisi nelle banche dati del genoma umano delle sequenze del DNA delle due regioni cromosomiche mappate (7q36 e 8q12) ha rivelato che esse includono più geni e che le proteine codificate da alcuni di questi geni hanno attività e funzioni cellulari che possono avere un ruolo nella determinazione dell'orientamento sessuale maschile. Pertanto questi geni individuati come candidati posizionali ora sono considerati anche candidati funzionali "deboli" per tale ruolo.
Nella regione cromosomica Chr7q36 mappa il gene del "recettore di tipo 2 del peptide intestinale vasoattivo (VIP)" che attiva l'enzima adenilato ciclasi in risposta al VIP ed il VIP funziona come un ormone neuroendocrino. Il recettore di tipo 2 è essenziale per lo sviluppo del nucleo ipotalamico soprachiasmatico e gli autori di questa ricerca considerano il recettore di tipo 2 un buon candidato

funzionale ad influenzare l'orientamento sessuale degli uomini perché gli uomini omosessuali hanno più sviluppato il nucleo ipotalamico soprachiasmatico.
La regione cromosomica 8p12 include il gene che codifica "l'ormone-1 che promuove il rilascio della gonadotropina" ed è localizzato in più nuclei dell'ipotalamo. Questo ormone stimola la sintesi ed il rilascio dell'ormone luteinizzante e dell'ormone follicolo stimolante che sono importanti regolatori della steroidogenesi nelle gonadi.
Adesso la ricerca è diretta ad identificare il meccanismo molecolare mediante il quale le proteine candidate funzionali contribuiscono a determinare l'orientamento sessuale maschile.

Alcune considerazioni sugli effetti delle mutazioni

Mutazioni silenti. Mutazioni puntiformi che modificano la sequenza nucleotidica di un codone; ma, per la degenerazione del codice, l'aminoacido codificato rimane lo stesso. Le mutazioni silenti sono dette anche isocodificanti perché non alterano la sequenza della proteina. Sono mutazioni silenti anche quelle che interessano gli introni o sequenze intergeniche. La degenerazione del codice interessa principalmente la terza base del codone ma anche la prima base dei codoni degli aminoacidi leucina ed arginina.
Mutazioni conservative. Mutazioni che causano la sostituzione di un aminoacido avente il residuo chimicamente simile a quello sostituito. Aminoacidi chimicamente simili (es. acido aspartico ed acido glutammico) hanno anche codoni simili. La sostituzione della terza base dei due codoni dell'acido aspartico porta alla formazione dei due codoni dell'acido glutammico. Egualmente la sostituzione della prima base del codone della leucina porta alla formazione del codone della valina; ambedue gli aminoacidi hanno residui idrofobici.
Le mutazioni conservative non alterano l'attività molecolare e la funzione cellulare della proteina codificata ma possono introdurre proprietà minori o subdole. Le proprietà minori e subdole si manifestano quando la normale fisiologia è portata a livelli limite (es. sforzo muscolare, digiuno prolungato, esposizione a luce intensa, ecc.) oppure perché la proteina mutata associa ed è sensibile a molecole contenute negli alimenti, presenti nell'ambiente (inquinanti e allergeni) oppure a farmaci ed anestetici.
Mutazioni conservative, fenotipi complessi e patologie complesse.
Le combinazioni diverse di alleli portatori di mutazioni conservative appartenenti a più geni responsabili dell'espressione di un dato fenotipo normale complesso contribuiscono alla variabilità normale di quel fenotipo complesso ed anche all'insorgenza di alterazioni del fenotipo normale definite come patologie complesse.
La funzione normale e l'alterazione patologica della stessa funzione dipendono da combinazioni diverse di alleli appartenenti allo stesso gruppo di geni. Differenze di combinazioni di alleli di uno stesso gruppo di geni sono

responsabili della variabilità dei caratteri normali complessi (esempio, variabilità dell'altezza corporea e della pressione arteriosa normale) e della variabilità dei sintomi di una stessa patologia complessa (esempio, variabilità dei valori della ipertensione arteriosa).

Mutazioni patologiche recessive. Le mutazioni patologiche recessive causano la patologia quando sono omozigoti. In eterozigosi, la ridotta quantità di proteina può essere sufficiente a mantenere la normale funzione cellulare perché alcune proteine hanno concentrazioni cellulari in eccesso rispetto alle esigenze funzionali della cellula (esempio, gli enzimi regolati solo dalla concentrazione del substrato). In alcuni casi la cellula compensa il deficit dell'attività molecolare con una maggiore sintesi della proteina e/o incrementando l'attività molecolare della proteina (figura D-6). Mutazioni di enzimi regolati solo dalla concentrazione del substrato che partecipano a vie metaboliche non di regolazione, in genere causano patologie recessive (es. fenilchetonuria, tirosinemia, alcaptonuria).

Mutazioni patologiche dominanti. Le mutazioni patologiche dominanti causano la patologia anche in condizioni di eterozigosi. La perdita, anche parziale, dell'attività molecolare della proteina mutata è sufficiente a causare la patologia perché la proteina è sintetizzata in quantità sufficienti per la normale fisiologia cellulare solo in omozigosi degli alleli normali, oppure perché la cellula non ha i meccanismi molecolari per compensare tale perdita. Mutazioni di enzimi regolati o partecipanti a vie di regolazione in genere sono dominanti perché gli enzimi regolati, essendo responsabili del "rate limiting step" (passo limitante la velocità), per poter regolare il flusso metabolico hanno attività catalitiche inferiori a quelle degli enzimi regolati solo da substrato facenti parte della stessa via metabolica. Pertanto è sufficiente una piccola perdita dell'attività molecolare della proteina codificata per alterare la regolazione ed instaurare la patologia (es. glucocinasi delle cellule-beta del pancreas nella patologia MODY2).

Una stessa patologia dominante può avere sintomi clinici diversi ed insorgere ad età molto diverse o non manifestarsi per tutta la vita del portatore della mutazione. Questa eterogeneità di espressione delle patologie dominanti è spiegata assumendo che la composizione degli alleli degli altri geni del portatore della mutazione dominante possa influenzare il manifestarsi della patologia ed a questa azione possono contribuire il tipo di alimentazione dell'individuo portatore della mutazione e/o l'ambiente nel quale ha vissuto.

Tutto questo suggerisce che anche le patologie monogeniche dipendano da alleli di più geni, anche se in forma molto limitata.

Si può quindi assumere che lo stato di salute e le patologie genetiche dipendano sempre da combinazioni di alleli di più geni.

Il dubbio è scomodo ma solo gli sciocchi non ne hanno.
Voltaire - François-Marie Arouet

APPENDICE A

Funzione del gene, attività molecolare e funzione fisiologica della proteina codificata.

La funzione del gene è la funzione che il prodotto del gene (mRNA, rRNA, tRNA o proteina) svolge nella fisiologia della cellula e dell'organismo. La funzione fisiologica della proteina è la funzione che la proteina svolge nella cellula dove è sintetizzata, nei fluidi biologici e/o in altre cellule. L'attività molecolare di una proteina è l'attività chimica e fisica della molecola proteica. I geni degli rRNA, tRNA e di altri RNA operano mediante gli RNA codificati e per queste macromolecole valgono le stesse considerazioni qui fatte per le proteine.
L'attività molecolare di una proteina viene anche chiamata attività biochimica (è ristretta alle attività chimiche) o attività biologica (definizione più generica: talvolta è usata per indicare complessivamente l'attività molecolare e la funzione fisiologica di una proteina).
Più raramente l'attività molecolare di una proteina viene chiamata funzione (senza l'aggettivo fisiologica), ad esempio si parla della (struttura e) funzione delle proteine intendendo la funzione molecolare, cioè l'attività molecolare che è analizzata con tecnologie biochimiche e/o biofisiche. Anche il ruolo che parti di macromolecole (es. sequenze promotrici di un gene, sequenze segnale di proteine, residui aminoacidi responsabili del legame del substrato e della catalisi) hanno nell'attività molecolare della macromolecola di cui fanno parte, è spesso indicato come funzione. Quindi è necessario considerare il contesto in cui questi termini sono inseriti per comprenderne il giusto significato.
Qui si tende a puntualizzare il significato diverso dei due termini (attività molecolare e funzione fisiologica) al fine di meglio comprendere le potenzialità ed i limiti delle tecnologie del DNA ricombinante perché con queste tecnologie si possono ottenere informazioni dirette sulla funzione di un gene e sulla funzione fisiologica della proteina da esso codificata, mentre forniscono informazioni indirette sull'attività molecolare delle proteine (capitolo 2). Per dimostrare l'attività molecolare delle proteine occorrono tecnologie biochimiche e biofisiche (es. dosaggio dell'attività catalitica di un enzima, analisi dei cambiamenti di conformazione di una proteina).
La differenza tra attività molecolare e funzione fisiologica di una proteina appare evidente quando uno stesso enzima o isoforme di uno stesso enzima aventi la stessa attività catalitica svolgono funzioni fisiologiche diverse. Ad esempio l'enzima lattico deidrogenasi (tetramero) delle cellule epatiche e renali ha la funzione fisiologica di favorire l'importazione dell'acido lattico dal sangue nelle cellule convertendolo in acido piruvico che poi viene utilizzato per sintetizzare glucosio (gluconeogenesi). Mentre l'isoforma muscolare dell'enzima

Sommario.

La funzione del gene si identifica con la funzione fisiologica del suo prodotto (RNA o proteina).
L'attività molecolare di una proteina è data dalle proprietà fisiche e chimiche che la proteina acquista assumendo la sua conformazione naturale (terziaria o quaternaria).
La funzione fisiologica di una proteina è data dalla funzione svolta nelle cellule (funzione cellulare) o nei liquidi biologici per azione della sua attività molecolare.

attività molecolare	-------->	funzione fisiologica
legare specificamente e reversibilmente una molecola nella cellula. mioglobina deposito di O_2 nella fibra	-------->	incremento della quantità della molecola (concentrazione) che può essere contenuta all'interno di una cellula. Esempio: muscolare.
legare specificamente e reversibilmente una molecola nei fluidi biologici.	-------->	incremento della quantità della molecola che può essere trasportata dal sangue. La funzione è diversa in relazione alla molecola trasportata. Esempi: l'emoglobina trasporta O_2 dai polmoni ai tessuti e l'albumina trasporta lipidi della dieta e quelli rilasciati dai tessuti.
catalisi enzimatica	-------->	capacità da parte della cellula di accelerare la sintesi o la degradazione di metaboliti. Funzione diversa in relazione alla reazione catalizzata (sintetica o degradativa), al metabolita sintetizzato o degradato, ed alla cellula in cui è localizzata la proteina (es. glucocinasi).
legare specificamente molecole segnale e trasmettere il segnale portato dalla molecola ad altre componenti cellulari meccanismo di trasduzione del segnale (via Ca) ed al tipo di cellula in cui è presente il	-------->	Il segnale proveniente da una cellula esocrina è trasmesso ad un'altra cellula da uno specifico recettore. Esempi: recettori degli ormoni e dei fattori di crescita. La funzione fisiologica è diversa in relazione alla molecola segnale, al AMPc, GMPc, inositolo-P, P-lipidi, ioni recettore.
Contrattilità	-------->	Contrazione della muscolatura. L'attività biologica è diversa in relazione al tipo di muscolo (movimento degli occhi, postura, movimento degli arti, respirazione, contrazioni del cuore, ecc.)

lattico deidrogenasi ha la funzione di rigenerare il coenzima NAD-ridotto (NADH) a NAD-ossidato (NAD) facendolo reagire con l'acido piruvico nella fibra muscolare che si contrae in condizioni di anaerobiosi. Ciò ha l'importante funzione di evitare il blocco della glicolisi per carenza di NAD ossidato. L'acido lattico è quindi esportato dalla fibra nel sangue ed è poi catturato dal fegato e dai reni che lo rigenerano a glucosio per riversarlo nel sangue (ciclo di Cori). L'enzima adenosina deaminasi che in alcune cellule ha la funzione di partecipare al catabolismo della adenosina, nei linfociti T, pur catalizzando la stessa reazione, svolge l'importante ruolo di permettere il normale sviluppo di queste cellule. La deficienza genetica di adenosina deaminasi causa una forma grave di immunodeficienza. Le diverse funzioni fisiologiche che l'enzima glucocinasi svolge nel fegato (metabolismo del glucosio) e nel pancreas (regolazione del rilascio di insulina) sono descritte nel capitolo 4
(patologia MODY2). Esistono anche proteine con attività molecolari e quindi anche funzioni fisiologiche diverse: monomeri di enzimi glicolitici del citoplasma (LDH e 3P-gliceraldeide deidrogenasi) hanno la funzione nucleare di fattori di trascrizione, l'enzima aconitasi (contenente ferro) ha un funzione cellulare nel metabolismo ed un’altra come proteina regolatrice della traduzione delle proteina ferritina e recettore della proteina transferrina (appendice B) e l'enzima citoplasmatico carbinolammina-deidratasi è anche localizzato nel nucleo ove svolge la funzione di fattore di trascrizione con il nome DCOH.

A Tenpo. A tenpo chi sa sa e chi non sa su' danno.
Francesco Omisi.

APPENDICE B

Cenni di regolazione dell'espressione genica negli eucarioti

<u>Condensazione e decondensazione della cromatina.</u>

Il DNA dei cromosomi degli eucarioti è associato ad ottameri $(H2A, H2B, H3 \text{ e } H4)_2$ costituiti da proteine basiche (istoni). Il DNA avvolto a spirale sugli ottameri forma i nucleosomi ed assume una struttura più compatta e più larga. Avvolgimenti successivi dei nucleosomi costituiscono delle strutture ancora più condensate (i solenoidi) e successivamente strutture ancora più compatte. La struttura più compatta è quella dei cromosomi metafasici dove il DNA, dati tutti gli avvolgimenti su se stesso, ha una lunghezza 10.000 volte più corta di quella che avrebbe se fosse in doppia elica distesa. La condensazione determina l'ispessimento dei cromosomi metafasici rendendoli visibili anche al microscopio ottico. Nelle altre fasi del ciclo cellulare la cromatina in certe zone è meno condensata (eucromatina, attiva in trascrizione o in replicazione) rispetto a quella condensata (eterocromatina, biologicamente inattiva). Le regioni di cromosoma che sono trascritte o replicate sono decondensate ma non completamente prive di nucleosomi, comunque espongono il DNA all'azione degli enzimi sintetici (DNA ed RNA polimerasi). La decondensazione dei cromosomi è associata alla perdita dell'istone H1, all'acetilazione ed ubiquitinazione degli istoni dei nucleosomi. L'atomo di N (azoto) del residuo degli aminoacidi basici (lisina ed arginina) è carico positivamente e forma legami salini con i gruppi di acido fosforico (carichi negativamente) del DNA. L'acetilazione degli atomi di N elimina la loro carica ed in questo modo favorisce il distacco/rilassamento dei nucleosomi dal DNA e quindi permette l'attacco dei fattori di trascrizione e delle polimerasi.

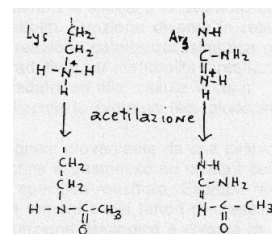
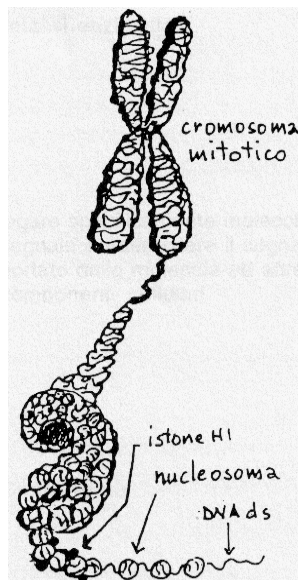



Figura B-1. a) Nei cromosomi metafasici il DNA in doppia elicaè reso più compatto di circa 10.000 volte. Il DNA del genoma diploide di ogni ingola cellula umana è costituito da circa 6 miliardi di coppie di basi, che allineate hanno una lunghezza di circa 2 metri. Un uomo di 70kg di peso è costituito da circa 10^{14}cellule il cui DNA ha in totale una lunghezza di 10 miliardi di Km (circa 1300 volte la distanza dalla terra al sole).
b) L'acetilazione dei residui di lisina e dì arginina degli istoni causa la perdita della carica dell'azoto.

<u>Metilazione e demetilazione del DNA a livello della citosina.</u>

La metilazione avviene a carico della citosina posta in sequenza con la guanina (CpG). Le coppie CpG sono ripetute e sparse nel genoma (capitolo 1) e la metilazione interessacitosine di CpG di tutti i cromosomi. Le regioni promotrici contenenti CpG metilate sono associate a geni inattivi. Si assume che la metilazione delle citosine inibisca l'associazione dei fattori di trascrizione e delle RNA-polimerasi. La metilazione favorisce lamutazione della citosina in timina per deaminazione ed ossidazione spontanea.

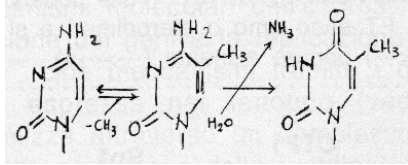



Figura B-2. Residui di citosina, citosina metilata e timina.

<u>Gli enzimi RNA-polimerasi degli eucarioti sono 3:</u>

RNA-polimerasi I	sintetizza rRNA precursore dei 18s, 28s e 5,8s RNA.
RNA-polimerasi II	sintetizza mRNA e piccoli RNA speciali.
RNA-polimerasi III	sintetizza tRNA, 5sRNA e piccoli RNA speciali.

Ognuna di queste RNA-polimerasi si lega a promotori strutturalmente diversi.

La RNA-polimerasi II ha scarsa affinità per i promotori dei geni e la sua associazione al DNA è permessa da proteine dette "fattori di trascrizione" che si legano a specifiche sequenze dette "sequenze promotrici". Alcune di queste sequenze sono comuni a più geni ed hanno tra loro similarità di sequenza (determinano delle sequenze consenso)(figura B-4).

<u>Fattori di trascrizione degli eucarioti</u>

I fattori di trascrizione (FT) degli eucarioti e procarioti sono proteine piccole (60-90 aminoacidi), hanno almeno tre domini (figura B-3) ciascuno dei quali con strutture proteiche specifiche raggruppati in pochi tipi.

<u>Domini di legame al DNA</u>

strutture: elica-giro-elica (pochi FT eucarioti, molti FT procarioti)
zinc finger (molti FT eucarioti, pochi FT procarioti)

Oltre a queste parti comuni a più FT, i domini di interazione con il DNA includono parti di sequenza specifiche per legarsi alla regione promotrice. Spesso l'interazione con le basi avviene mediante legami ad H di residui aminoacidici di Asn, Gln, Glu, Lys e Arg ed anche con legame idrofobico con il gruppo metile della timina. La maggior parte dei legami specifici si formano con le basi nucleotidiche nella scanalatura grande della doppia elica.

<u>Dominio di dimerizzazione</u>

strutture: leucina zipper (molti FT eucarioti, pochi FT procarioti)
elica-ansa-elica (eucarioti con regione basica che si lega al DNA)

<u>Domini di interazione con la RNA-polimerasi II</u>

strutture: ricco di glutamina (Q)
ricco di prolina (P)
acido, ricco di aspartato (D) o glutammato (E).

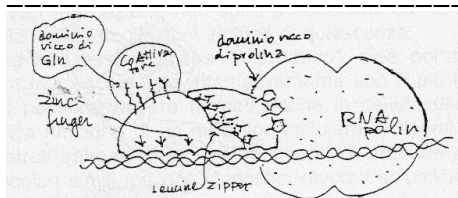



Figura B-3. Il disegno rappresenta un FT ipotetico.

In genere i FT che si legano al DNA hanno affinità per qualsiasi sequenza nucleotidica di DNAds (Kaff≈10^{6-8}), ma per le sequenze verso le quali hanno specificità di legame, l'affinità aumenta di 10^{5-7} volte portando la Kaff a 10^{13-15}. Molti FT sono omo o eterodimeri e si legano a sequenze promotrici palindromiche (---><---) oppure allineate (--->--->).

Figura B-4. Alcune sequenze promotrici e relativi FT comuni a più geni. I FT sono rappresentati come monomeri e le diverse ombreggiature indicano domini diversi di uno stesso monomero. P = residuo di prolina, Q = residuo di glutamina.

Le sequenze promotrici sono indicate per convenzione con la sola sequenza del filamento 5'--3', in realtà i FT si legano al doppio filamento di DNAds. La disposizione delle sequenze non permette il contatto tra i FT e la RNA-polimerasi. Ciò è realizzato mediante ripiegamenti del DNA e l'intervento di altre proteine dette coattivatori (CoA) che fanno da ponte tra i FT e la RNA-polimerasi. Per alcuni geni si ha azione di regolazione (attivazione o inibizione) anche da parte di istoni dei nucleosomi (figura B-5).

Figura B-5. Il disegno rappresenta un ipotetico complesso di inizio di trascrizione.

Meccanismo genetico di base che dal gene porta alla formazione della proteina

La biologia molecolare, definendo i meccanismi molecolari dell'espressione genica, ha portato a considerare l'espressione dei geni e la replicazione del DNA come "meccanismi genetici di base", cioè meccanismi tramite i quali l'informazione genetica è replicata ed espressa nel fenotipo (vedere

prefazione). L'espressione dei geni si realizza attraverso un complesso ed ordinato insieme di reazioni che portano alla produzione della proteina(e) da esso codificata (figura B-6). Questo processo per realizzarsi richiede l'intervento di quasi tutti i processi metabolici cellulari che includono la produzione di nucleotidi trifosfati, acidi ribonucleici, proteine con la formazione di complessi sopramolecolari (ribosomi, poliribosomi, membrane).
Negli eucarioti (uomo incluso), le sequenze codificanti dei geni (esclusi alcuni geni degli istoni) sono divise in segmenti (esoni) separati da segmenti di sequenze non codificanti (introni). Questa struttura in esoni ed introni si ritrova nel trascritto primario del gene ed attraverso un processo di maturazione gli esoni sono separati dagli introni che vengono eliminati (degradati per azione di enzimi RNA-asi), mentre gli esoni si uniscono (splicing) per formare la sequenza continua codificante la proteina. Il taglio nel punto di confine tra esoni ed introni è indicato da particolari sequenze consenso poste al 5' e 3' dell'introne che al 5' iniziano con GT ed al 3' terminano con AG (regola AG-GT, figura B-6). All'interno degli introni esistono altre sequenze che rendono possibile e specifico lo splicing. La maturazione del mRNA include anche: -la formazione del cap (cappello) del mRNA, cioè la legatura covalente di una 7-metilguanosina al primo nucleotide del trascritto primario mediante un legame 5'-5' fosfodiestere; -l'aggiunta (per reazioni successive catalizzate dell'enzima poli(A) polimerasi) di un segmento di poli-A (circa 200 A) al mRNA. Il trascritto primario viene tagliato al 3' a distanza di circa 15-30 basi di una sequenza segnale (AAUAAA). Dopo il taglio viene aggiunto il poli-A che non è codificato nel gene. Le funzioni attribuite al cap includono: proteggere l'mRNA dalle esonucleasi, facilitarne lo splicing, il trasporto dal nucleo al citoplasma e l'associazione alla subunità 40s del ribosoma. Le funzioni del poli-A sono simili a quelle del cap: stabilizzare l'mRNA, facilitare il suo trasporto dal nucleo al citoplasma e l'associazione ottimale con i ribosomi durante la traduzione. Gli mRNA degli istoni varianti di replicazione non sono poliadenilati ed hanno vita più breve degli mRNA poliadenilati degli istoni varianti di quiescenza.
Il meccanismo di splicing può essere anche "alternativo", cioè portare alla formazione di più tipi di mRNA escludendo alternativamente uno o più esoni. I due o più mRNA diversi possono essere formati anche in cellule diverse e contribuire ad una specifica funzione di un dato tipo cellulare. I diversi mRNA codificheranno proteine strutturalmente diverse (mancanti di uno/due domini) che avranno attività molecolari simili ma non identiche: diversa solubilità al finedi poter essere solubili ed attive in tipi cellulari diversi; potranno avere o non avere siti di associazione di molecole regolatrici; avere o non avere domini di associazione della forma monomerica e quindi formare oligomeri o esclusivamente monomeri (le struttura oligomerica è necessaria per poter operare l'effetto di regolazione cooperativo); avere o non avere siti di legame covalente per molecole (modificazioni post-traduzionali) che conferiscono alla proteina maggiore solubilità, possibilità di associarsi a membrane, regolazione da parte di ormoni o fattori di crescita attraverso reazioni di

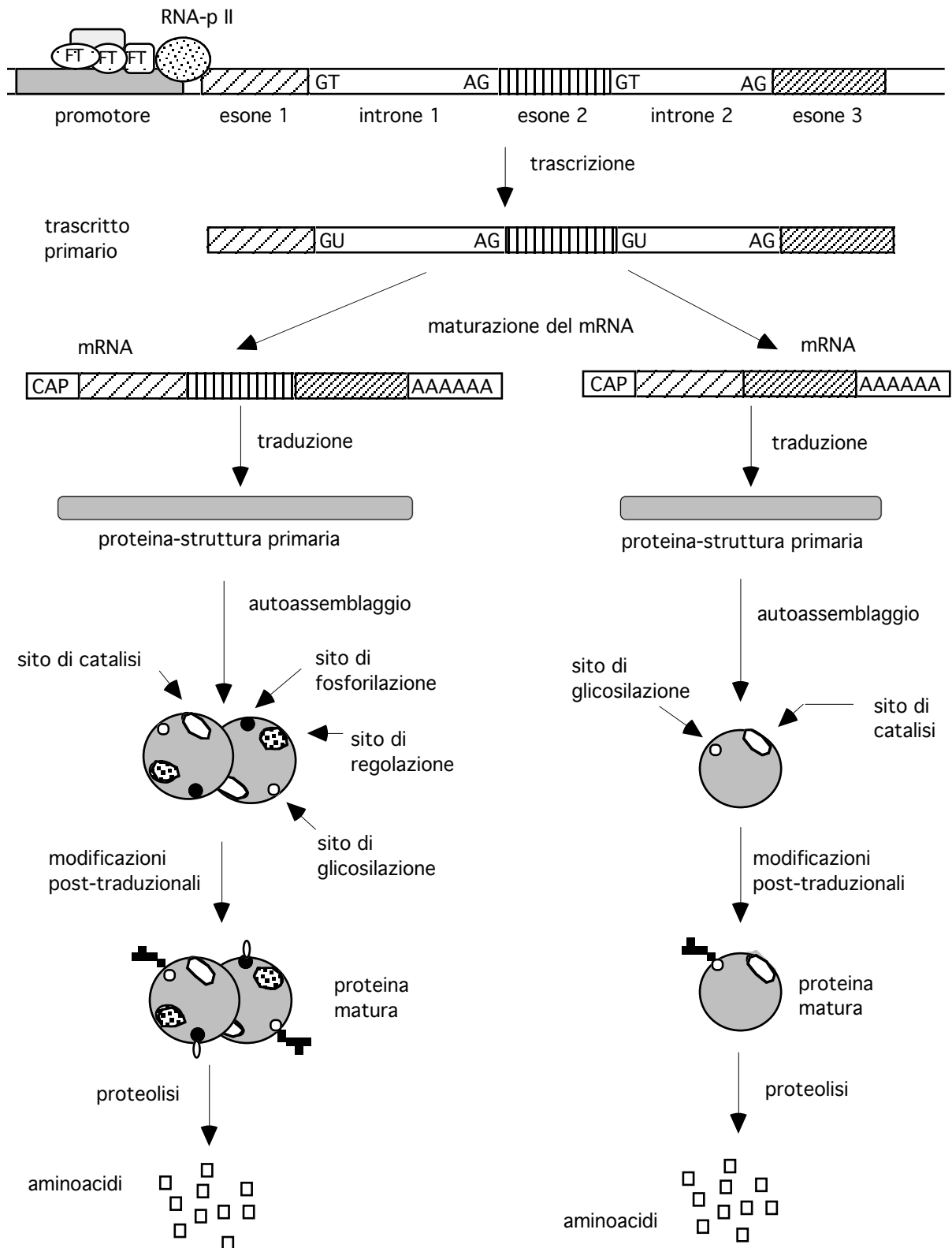



Figura B-6. Schema del meccanismo genetico di base che dal gene porta alla formazione delle proteine (fenotipo). Lo splicing alternativo porta alla formazione di due proteine: una monomerica e l'altra dimerica dotata di regolazione da effettore e da fosforilazione.

fosforilazione/defosforilazione; essere più o meno sensibili ai processi diproteolisi cellulare e quindi avere nella cellula una vita più o meno lunga. Risulta evidente che ogni singolo passaggio dell'espressione genica è reso specifico dal riconoscimento molecolare di particolari strutture: sequenze nucleotidiche del DNA o RNA da parte di acidi nucleici e/o proteine. Ad esempio: per l'inizio e fine della trascrizione, per lo splicing, l'introduzione del cap e per la poliadenilazione del mRNA, l'inizio e lo stop corretto della traduzione. Il polipeptide sintetizzato ha nella sua sequenza aminoacidica (struttura primaria) tutte le informazioni per assumere la sua struttura terziaria e quaternaria formando siti catalitici, di associazione specifica o di legame covalente con altre molecole, di sensibilità alla proteolisi. Tutte queste sequenze e strutture che interagiscono specificamente tra loro sono luoghi di regolazione al fine di avere la proteina più o meno concentrata o più o meno attiva per le necessità delle cellule e dell'organismo.

Regolazione della traduzione negli eucarioti

Il controllo dell'espressione genica nei procarioti è principalmente a livello della trascrizione e ciò è reso possibile dalla vita breve (media = 1,5 minuti) dei loro mRNA. Se il gene non viene trascritto, l'mRNA viene rapidamente degradato e la traduzione si interrompe. Quindi la trascrizione regola direttamente la sintesi proteica perché regola la concentrazione del mRNA.
Negli eucarioti sintesi e traduzione del mRNA avvengono in compartimenti subcellulari diversi (nucleo e citoplasma) e globalmente l'mRNA di una cellula ha una vita media di circa 3h (con estremi da pochi minuti ad alcuni mesi) per cui un arresto della trascrizione non produce effetti rapidi sulla concentrazione del mRNA e quindi sulla sintesi delle proteine. Negli eucarioti oltre al controllo della trascrizione esiste un controllo della traduzione del mRNA. Il controllo della trascrizione rimane primario perché senza mRNA non ci può essere controllo della traduzione, ma anche la traduzione è regolata positivamente e negativamente mediante specifici meccanismi molecolari. Il controllo della traduzione esiste anche nei procarioti ma, data la labilità degli mRNA, è un tipo di regolazione meno esteso rispetto agli eucarioti.
Un esempio di regolazione della traduzione è dato dalla regolazione della sintesi della ferritina e del recettore della transferrina. Lo ione Fe^{3+} è trasportato nel sangue dalla proteina transferrina. Il complesso transferrina-Fe^{3+}, mediante endocitosi, penetra nelle cellule (es. epatociti) che contengono i recettori specifici per la transferrina. Il Fe^{3+} è rilasciato nel citoplasma e si lega ad un'altra proteina (ferritina), che è la forma di deposito di Fe^{3+} intracellulare. Se la concentrazione del [Fe^{3+} libero] aumenta nel citoplasma (se eccessiva può essere tossica) essa provoca un incremento della traduzione del mRNA di un'altra proteina, la ferritina, e un incremento della distruzione del mRNA del recettore della transferrina. Le due azioni determinano rispettivamente l'aumento di sintesi di ferritina (incrementa l'accettore di Fe^{3+}) e la riduzione di sintesi di recettori della transferrina (riduzione di entrata nella cellula di Fe^{3+})

che risultano in un decremento di Fe^{3+} libero. Questa funzione è realizzata da strutture simili a forcine per capelli dette "gambo-cappio" (stem-loop) aventi sequenze simili, presenti nelle due specie di mRNA e da una proteina regolatrice che si lega ai gambo-cappio ed ha attività diverse nelle due specie di mRNA. L'mRNA-ferritina ha un gambo-cappio al 5' a cui si lega la proteina regolatrice e così legata inibisce la traduzione della ferritina. L'mRNA del recettore della transferrina ha al 3' cinque gambo-cappio a cui si lega la proteina di regolazione che in questo modo protegge l'mRNA dall'attacco delle RNAasi. Quando la [Fe^{3+}] aumenta nel citoplasma, la proteina regolatrice associa Fe^{3+}, cambia conformazione e si dissocia da ambedue le specie di mRNA. Ora l'mRNA-ferritina è libero di legarsi ai ribosomi e di dirigere la sintesi di ferritina mentre l'mRNA dei recettori della transferrina viene più facilmente attaccato dalle RNAasi ed in questo modo si riduce la sintesi dei recettori della transferrina. Poiché tutte i componenti cellulari (eccetto il DNA) hanno un tempo di vita, la ridotta o nulla sintesi dei recettori porta alla riduzione della loro concentrazione. Il meccanismo è reversibile: quando la [Fe^{3+}] diminuisce nel citoplasma, il complesso Fe^{3+}-proteina di regolazione si dissocia e la proteina si associa ai gambo-cappio determinando l'inibizione della sintesi di ferritina e l'incremento della sintesi dei recettori della transferrina.

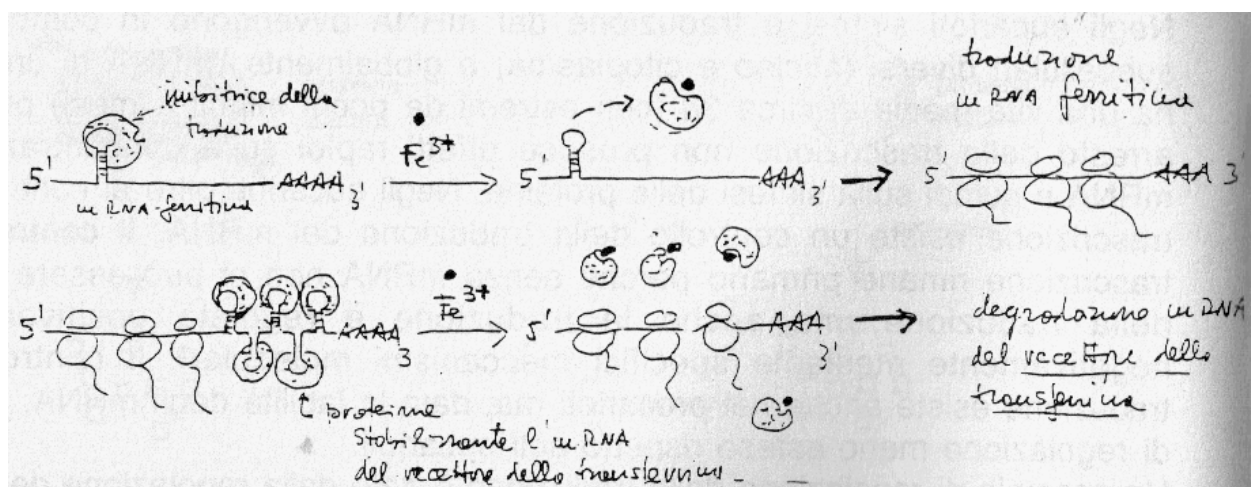
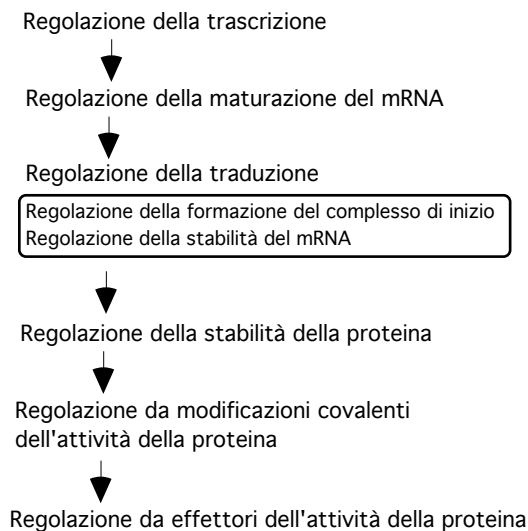
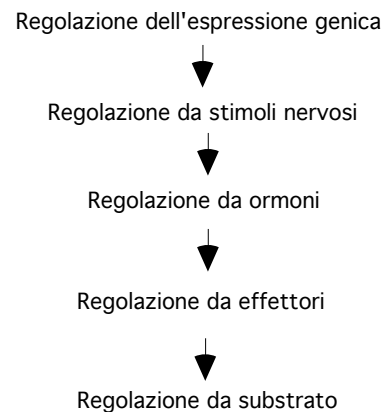



Figura B-7. Regolazione a livello della traduzione della ferritina e del recettore della transferrina.

La proteina regolatrice della ferritina è identica all'enzima aconitasi che ha come cofattore il ferro. Questo è un caso particolare in cui una stessa proteina ha due diverse attività molecolari (catalisi ed associazione al mRNA) e due diverse funzioni fisiologiche (metabolica e di regolazione della traduzione).
Oltre che dalla regolazione della traduzione, la concentrazione cellulare di una proteina è regolata da meccanismi che regolano la sua distruzione. Tutte le proteine hanno un proprio tempo di vita, anche se l'attività di alcune proteine deve essere sempre presente nella cellula. Si ha una continua sintesi e degradazione che determina un continuo <u>ricambio</u> (turnover) della proteina. Il turnover è in relazione alla regolazione delle funzioni cellulari dei tessuti e degli

organi e ai periodi di digiuno (sono degradate le proteine di tutti gli organi ma soprattutto quelle epatiche e muscolari ed i loro aminoacidi utilizzati per sostenere la gluconeogenesi).
L'attività molecolare della proteina, oltre che dalla regolazione della sua concentrazione, è regolata dall'associazione specifica con altre molecole (es. il substrato per gli enzimi, la regolazione da substrato, il recettore con il rispettivo ormone), da reazioni covalenti (es. fosforilazione e defosforilazione) e da reazioni di associazione con effettori (regolazione allosterica). Molta dell'energia (ATP e di altri NTP) prodotta dalla combustione degli alimenti è utilizzata per mantenere i sistemi molecolari di regolazione che continuamente sintetizzano e distruggono i propri elementi al fine di poter agire prontamente e prontamente inibire la propria azione. Esiste una gerarchia dei sistemi di regolazione dell'espressione genica e della regolazione delle proteine, dove la regolazione della trascrizione è primaria essendo la prima ad intervenire.

Educare una persona nella mente e non nella morale, è educare una minaccia alla società.
Theodore Roosevelt

Appendice C

Incremento della specificità del riconoscimento molecolare[18, 29-42]

Il riconoscimento molecolare è l'associazione specifica tra molecole mediante interazioni deboli (legami a H, legame salino, legame idrofobico, forze di van der Waals e legami di coordinazione) che porta alla formazione di complessi DNAss-DNAss, DNAss-RNA, DNA-NTP, RNA-NTP e complessi tra proteina (P) e legante (L). Il legante delle proteine può essere una piccola molecola (es. metabolita), un glucide polimerico (es. glicogeno), un lipide sopramolecolare (es. membrana cellulare), un'altra proteina o un acido nucleico (DNA o RNA).
Il riconoscimento molecolare è la prima fase di ogni attività molecolare degli acidi nucleici e delle proteine. Ad esempio: prima si forma il complesso <singolo NTP>-DNAss poi l'enzima DNA-polimerasi-II catalizza il legame fosfodiestere, prima si forma il <complesso fattore di trascrizione>-<sequenza nucleotidica del promotore> poi il fattore di trascrizione attiva del RNA-polimerasi-II, prima si forma il complesso enzima-substrato poi avviene la catalisi, prima si forma il complesso ormone-recettore poi si ha la trasduzione del segnale.
La specificità del riconoscimento molecolare tra polinucleotide e polinucleotide complementare e tra P ed L è determinata dalla complementarità di forma delle superfici di contatto (come una medaglia con il proprio calco) e la complementarità di carica (gruppi atomici carichi positivamente che interagiscono con gruppi carichi negativamente e gruppi idrofobici con altri gruppi idrofobici) esistenti tra le molecole che si associano spontaneamente spinte dall'agitazione molecolare provocata dal calore. La precisa complementarità di forma delle molecole che si associano, è necessaria per poter formare più interazioni deboli tra le loro superfici di contatto, perché la forza delle interazioni deboli dipende dalla distanza tra gli atomi che interagiscono (troppo vicini si respingono, lontani non hanno forza di legame). Pertanto per poter avere una superficie di interazione sufficientemente grande da essere capace di un numero di interazioni deboli sufficiente a formare un complesso stabile, almeno una delle due molecole interagenti deve essere una macromolecola. Gli ioni metallici possono complessarsi anche con molecole non macromolecolari, formando con esse dei legami di coordinazione.
Il grado di specificità di formazione del complesso è in relazione alla disposizione ed al numero delle interazioni deboli che si formano tra le molecole associate, cioè ai punti di interazione tra le due superfici. Se le superfici di contatto tra le molecole interagenti non avessero una forma complementare sufficientemente estesa non si formerebbe un numero di interazioni sufficienti a dare un alto grado di specificità al riconoscimento molecolare ed a mantenere le molecole associate.
Il grado di affinità tra le molecole interagenti (indicata anche come spontaneità a formare il complesso), è definito dal valore di più tipi di parametro: valore

dell'energia libera liberata nel formare i legami deboli, valore della costante di affinità (Ka) che è uguale al rapporto delle costanti cinetiche di reazione k_1/k_2 ed al rapporto della concentrazione del complesso formato diviso per il prodotto delle concentrazioni delle molecole interagenti (che in genere sono due) con la reazione in condizioni di equilibrio. Il valore della Ka è in relazione inversa al valore della variazione di energia libera liberatata nel formare il complesso e quest'ultimo valore da la misura della stabilità del complesso, cioè il valore dell'energia libera necessaria per rompere le interazioni deboli e dissociare il complesso. Il valore di energia libera delle interazioni deboli è mediamente circa 1kcaloria/mole, pertanto se il complesso è formato da 7 legami deboli (es. complesso emoglobina-3Pglicerato), l'energia libera liberata per formarli è -7kcal/mole che corrisponde ad una Ka di circa $10M^{-1}$; se si formano 10 legami deboli (es. complesso proteina Ras-GTP) si liberano -10kcal/mole e la Ka è circa 10^8M^{-1}.
La formazione dei complessi ha un valore di energia di attivazione molto basso che è facilmente superato alle temperature di 37°C. A causa della trascurabile energia di attivazione di formazione del complesso, è sufficiente che le molecole biologiche reagenti vengano in contatto con il giusto orientamento ed anche se hanno bassa energia, riescono a formare il complesso. E' sufficiente che le molecole incontrandosi "sguscino" via dall'acqua (rompano i legami a H con essa e/o con altri soluti) per formare il complesso. Pertanto nei fluidi biologici dato l'alto valore di specificità ed affinità delle molecole, ognuna di esse si trova quasi tutta in forma complessata. Il complesso è in concentrazione allo stato stazionario, cioè si forma e si rompe continuamente per permettere, quando necessario, una veloce utilizzazione delle molecole. Mancando la barriera data dall'energia di attivazione, l'energia per dissociare il complesso ha lo stesso valore dell'energia liberata per formarlo.
Queste caratteristiche differenziano le reazioni di formazione dei complessi dalle reazioni covalenti. Le reazioni covalenti hanno alti valori di energia di attivazione che costituiscono una alta barriera all'avvenire della reazione in ambedue i sensi della reazione. Pertanto nelle cellule possono esistere per tempi molto lunghi molecole che per la loro alta spontaneità dovrebbero essere convertite in prodotti di più basso valore energetico (es. glucosio ed O_2). La cellula controlla le reazione covalenti, praticamente ferme per la fisiologia cellulare, mediante la regolazione degli enzimi. La catalisi enzimatica incrementa di 10^6-10^8 la velocità delle reazioni covalenti portandola a valori molto vicini a quelli delle reazioni di associazione.
Nelle molecole biologiche esiste una specializzazione di legami: i legami covalenti sono i legami che tengono uniti gli atomi delle molecole e delle macromolecole, le interazioni deboli sono i legami che tengono uniti i complessi a cui partecipano le macromolecole. Le reazione covalenti catalizzate interessano soprattutto il metabolismo degli alimenti, quello energetico e la sintesi delle molecole e macromolecole. Le reazioni mediante interazioni deboli interessano soprattutto il riconoscimento molecolare, le molecole segnale e le vie di regolazione.

Le reazioni di formazione dei complessi biologici, data la bassa energia di attivazione, hanno costanti cinetiche di formazione (k_1) di alto valore, tra 10^6-$10^8 M^{-1}s^{-1}$) mentre la costante cinetica di dissociazione (k_2) ha valori molto più bassi che sono in ragione inversa al grado di stabilità del complesso. La stabilità dei complessi è in relazione alle attività delle macromolecole, infatti il grado di stabilità del complesso è determinante affinché la macromolecola possa svolgere la sua attività e quindi la sua funzione cellulare.
Ad esempio, alcuni enzimi, in relazione alla complessità della reazione catalizzata, hanno un relativamente alto valore della costante cinetica k_2. L'enzima piruvato carbossilasi ha k_1=4,5x$10^6 M^{-1}s^{-1}$ e k_2=2,1x$10^4 s^{-1}$, l'enzima ribonucleasi ha k_1=7,8x$10^7 M^{-1}s^{-1}$ e k_2=1,1x$10^4 s^{-1}$, pertanto questi complessi enzima substrato hanno una bassa stabilità e la reazione di formazione dei complessi enzima-substrato è altamente reversibile. Gli enzimi sono congegni molecolari che hanno micro-cambiamenti di conformazione durante la catalisi. Se il substrato formasse molte interazioni deboli, il complesso sarebbe più stabile e ciò potrebbe impedire il cambiamento di conformazione dell'enzima e quindi la catalisi, come avviene per certi inibitori competitivi più affini all'enzima del substrato naturale. Le proteine di deposito hanno valori di k_2 relativamente bassi. Ad esempio la mioglobina (Mb) ha k_1=$10^7 M^1 s^{-1}$ e k_2=10s^{-1}, il basso valore di k_2 determina un equilibrio più spostato a destra e quindi un'alta concentrazione di complessi Mb-O_2. Ciò rende la Mb capace di svolgere la funzione di immagazzinare nelle cellule una grande quantità di O_2, legante poco solubile in acqua; tuttavia essendo il legame tra Mb e O_2 una interazione debole (un singolo legame di coordinazione) la reazione di formazione del complesso Mb-O_2 è altamente reversibile. Quando la concentrazione dell'O_2 libero diminuisce perché utilizzato per la contrazione della fibra muscolare, il complesso Mb-O_2 si dissocia e la Mb svolge la sua funzione fisiologica di deposito aperto: immagazzinatrice e distributrice di O_2.
Nella formazione dei complessi biologici almeno uno dei reagenti è una macromolecola perché le macromolecole hanno dimensioni molecolari tali da poter avere sulla loro superficie un'area (spesso una cavità) sufficientemente grande da poter formare con il legante un numero di interazioni deboli capace di conferire un'alta specificità e stabilità al complesso.
In genere la specificità a formare il complesso aumenta con la stabilità del complesso stesso perché la stabilità del complesso è data dalla somma delle energie libere delle singole interazioni deboli e il numero delle singole interazioni è in relazione diretta con la specificità del complesso. Tuttavia in alcuni complessi si può avere una dissociazione tra affinità e specificità come nelle proteine allosteriche dove l'affinità verso il legante può essere variata cambiando la conformazione della proteina (es. emoglobina), mentre la specificità della reazione rimane la stessa. L'emoglobina durante il trasporto dell'O_2 cambia affinità per l'O_2 ma non la specificità per la stessa molecola. In altri casi, il complesso è formato da un certo numero di interazioni deboli, esempio 8, che conferiscono un buon grado di specificità, ma aventi forza inferiore a 1kC/mole, perché le distanze tra gli atomi che interagiscono sono

superiori o inferiori a quelle ottimali. In questo caso la specificità risulta alta come in altri complessi stabilizzati da 8 interazioni, ma l'affinità è inferiore (es. 5kC/mole di complessi invece delle 8kc/mole di complessi) data la minore forza dei legami.
Riassumendo le basi fisico-chimiche del riconoscimento molecolare sono le complementarità di carica e di forma delle superfici di contatto tra le molecole ed il calore. La complementarità di carica e di forma determina la formazione delle interazioni deboli una volta che le molecole siano venute in contatto.
Il calore provocando l'agitazione delle molecole nelle soluzioni biologiche svolge la doppia funzione costruttiva e distruttiva dei complessi. Determinando l'incontro delle molecole, il calore favorisce la formazione dei complessi e scuotendo i complessi ne provoca la dissociazione, rendendo facilmente reversibili reazioni anche se fortemente spostate verso la formazione dei complessi, anche se la reazione di dissociazione dei complessi è molto più lenta (tra 100 volte e miliardi di volte) di quella di formazione. L'energia cinetica fornita dal calore è sufficiente per dissociare anche complessi molto stabili (es. $Mb\text{-}O_2$). L'alto grado di reversibilità delle reazioni di formazione dei complessi è un fattore importante per determinare l'alta specificità del riconoscimento molecolare. Un esempio è dato dalle reazioni di ibridazione degli acidi nucleici fatte alla Tm proprio per permettere un'alta reversibilità della reazione al fine di ottenere la formazione specifica di complessi di DNAds. Durante la reazione di ibridazione si possono associare filamenti di DNAds non perfettamente complementari (es. diversi per una singola base) dotati anche di stabilità, tuttavia la continua associazione e dissociazione del DNAds provocata dal calore finirà con eliminare i complessi di DNAds meno stabili ed a favorire la formazione di quelli più stabili, fino ad avere solo complessi di DNAds costituiti da filamenti di DNAss perfettamente complementari. Il calore alla Tm porta ad avere una concentrazione pari al 50% dei complessi possibili di DNAds esattamente complementari, che sono in equilibrio con i filamenti di DNAss liberi.
Sebbene la specificità delle reazioni di associazione dei complessi biologici sia molto alta in alcuni casi essa viene incrementata dall'intervento di altre proteine. L'incremento di specificità può risultare dalla preventiva inattivazione di leganti affini alla proteina prima che essi possano legarsi ad essa al fine di evitare una attività molecolare non fisiologica. Un esempio di questo tipo è l'azione dell'enzima 11β-idrossisteroide deidrogenasi che nelle cellule del nefrone distale continuamente converte il cortisolo in cortisone, steroide non affine al recettore dei mineralcorticoidi (MR), ed in questo modo permette solamente all'aldosterone di legarsi al MR. Ci si può domandare perché la natura non abbia fatto le cose un poco più precise tra aldosterone ed MR. Una spiegazione può essere che l'enzima 11β-HSD selezionato nell'evoluzione per avere la funzione di controllare la concentrazione intracellulare del cortisolo ha finito poi per avere anche la funzione di incrementare la specificità del recettore MR ed essendoci l'enzima 11β-HSD, l'evoluzione non ha favorito incrementi di specificità tra MR ed aldosterone.

Un altro tipo di incremento della specificità del riconoscimento molecolare presente in natura è l'eliminazione degli errori di associazione dopo che essi si siano verificati.
Un esempio di questo tipo è dato dalla reazione di associazione tra codone ed anticodone del secondo e dei seguenti aminoacil-tRNA (aa-tRNA) nel sito A (aminoacidico) del ribosoma. La reazione di associazione avviene in una cavità del complesso <subunità grande>-<subunità piccola> del ribosoma ed è seguita dalla reazione di formazione del legame peptidico (punto da cui non si può tornare indietro in caso di inserimento di un aa-tRNA sbagliato e quindi di un aminoacido sbagliato nel polipeptide). Le similarità tra gli anticodici del tRNA che anti-codificano aminoacidi diversi offrono la possibilità di formazione di qualche legame ad H con codoni sbagliati, la possibilità di distorsioni transitorie del tRNA nella regione dell'anticodone ed il ridotto grado di libertà di diffusione del aa-tRNA nel sito A permettono ad un aa-tRNA sbagliato di rimanere nel sito un brevissimo tempo, tuttavia sufficientemente lungo a permettere la formazione del legame peptidico. Ciò accadrebbe se l'aa-tRNA non fosse legato ad una proteina, il fattore di elongazione EF (EF-Tu dei batteri e eEF1α negli eucarioti) che a sua volta lega GTP e sul GTP ha attività di idrolisi del fosfato in γ del GTP (GTPasica). Il complesso aa-tRNA-EF-GTP si inserisce nel sito A e la presenza di EF-GTP impedisce l'avvenire della reazione di formazione del legame peptidico per circa 1 millesimo di secondo (il tempo di idrolisi del GTP a GDP), il complesso aa-tRNA-EF-GDP permane per un altro millesimo di secondo, dopo di che il complesso EF-GDP si dissocia dal aa-tRNA il quale è ora libero di reagire per formare il legame peptidico. Il complesso EF-GTP, mantenendo per circa due millisecondi l'aa-tRNA non reattivo (per la formazione del legame peptidico), concede un tempo sufficientemente lungo a permettere la dissociazione degli aa-tRNA sbagliati. Gli aa-tRNA sbagliati, formando meno legami con il codone, sono complessati meno stabilmente con l'mRNA e si dissociano prima degli aa-tRNA giusti. I due millisecondi sono un tempo sufficientemente breve da evitare la dissociazione dei aa-tRNA corretti. A conferma di ciò l'osservazione che se si usa GTPγS (GTP con un atomo di S, zolfo, al posto di uno di P in gamma), EF idrolizza più lentamente, la traduzione è più fedele ma più lenta perché il tempo di non reattività del aa-tRNA è più lungo. Quindi l'intervento di una proteina cellulare (EF) migliora la specificità di associazione codone-anticodone. La scelta della velocità di idrolisi del GTP (operata da EF) è un compromesso che la Natura ha fatto per avere una buona fedeltà associata ad un buona velocità di sintesi proteica. Il compromesso non causa danni perché gli errori sono relativamente pochi ed una proteina con un aminoacido errato in genere non crea danni alla cellula perché molte copie della stessa proteina sono sintetizzate correttamente. Le proteine che legano GTP ed hanno attività GTPasica sono coinvolte in molti processi biologici importanti (es. trasmissione del segnale ormonale all'interno della cellula), esse hanno come per EF sopra indicato la funzione di <u>conta millisecondi</u>. Idrolizzato il GTP a GDP si dissociano da esso e perdono la loro attività molecolare, e per riacquistarla devono associare nuovo GTP.

Un altro esempio di incremento della specificità del riconoscimento molecolare è dato dal sistema di eliminazione dei nucleotidi non correttamente accoppiati (proofreading) durante la sintesi del DNA posseduto dall'enzima DNA-polimerasi-III. Il riconoscimento molecolare avviene tra la base del nucleotide del DNA stampo ed il NTP ad esso complementare, l'enzima DNA-polimerasi-III ha la funzione di catalizzare la reazione di formazione del legame fosfodiestere tra il NTP neoinserito e quello ad esso adiacente al 5'.
L'enzima DNA-polimerasi-III di E. coli, enzima della replicazione del DNA, ha una velocità di catalisi di 250-1000 nucleotidi al secondo e con questa velocità il riconoscimento molecolare tra le basi compie un errore (una base non correttamente accoppiata) per ogni 10.000-100.000 nucleotidi inseriti correttamente. Questa frequenza di errore viene ridotta di 100-1000 volte dalla presenza sullo stesso enzima dell'attività esonucleasica 3'-->5' (detta di correzione delle bozze, proofreading) per cui risulta che la DNA-polimerasi-III duplica il DNA commettendo un errore ogni 1-100 milioni di nucleotidi inseriti correttamente. L'accuratezza della replicazione risulta ancora maggiore perché i nucleotidi mal accoppiati (sfuggiti all'attività esonucleasica 3'-->5') sono corretti dal sistema di riparazione che riconosce i nucleotidi disaccoppiati (mismatch repair). Se la velocità di catalisi della DNA-polimerasi-III fosse più bassa gli errori sarebbero inferiori. Forse dovendo comunque correggerli, la Natura ha preferito una velocità maggiore seguita da una correzione degli errori.
Nonostante la correzione, basi errate permangono nel DNA replicato ed esse sono la maggiore causa di patologie genetiche trasmesse e somatiche. La fedeltà di replicazione nelle cellule dell'uomo è estremamente alta: per una sostituzione di una coppia di basi ogni 10^9-10^{10} solo 1-6 basi risultano errate durante la replicazione di 6 miliardi di basi del DNA umano.

L'ordine è economia di tempo e di spazio e di potente ausilio alla memoria.
Anonimo

Appendice D

Eterogeneità genetica della specie umana

Ogni individuo della specie umana (esclusi i gemelli monozigoti) nasce e rimane geneticamente diverso dagli altri individui (familiari inclusi), per cui le risposte ad agenti, fisici, chimici e biologici (farmaci inclusi) possono essere diverse da individuo ad individuo.

Il corredo cromosomico dell'uomo è diploide. La diploidia è una condizione necessaria per poter realizzare alcuni dei meccanismi responsabili dell'eterogeneità genetica ed inoltre essa rappresenta una forma di difesa contro le mutazioni dei geni che causano alterazioni nelle funzioni cellulari. Se un allele di un gene subisce una mutazione recessiva che altera l'attività molecolare della proteina espressa, l'altro allele dello stesso gene (sul cromosoma omologo), essendo integro, esprime una proteina normale che può mantenere la normale fisiologia cellulare. Ciò non si verifica quando la mutazione è dominante o dominante negativa.

Può accadere che la mutazione conferisca alla proteina una nuova attività molecolare a scapito di quella (o di una di quelle) che aveva. In questo modo l'organismo diploide possiede ambedue le funzioni. Ciò non sarebbe possibile se il corredo cromosomico fosse aploide. Inoltre quando due mutazioni favorevoli avvengono su geni diversi (loci diversi) appartenenti ad individui diversi di una stessa specie, con la riproduzione sessuata i due geni mutati possono ritrovarsi sullo stesso individuo che avrà così un doppio vantaggio selettivo. L'amore vince (o meglio la sessualità), infatti questo meccanismo non può realizzarsi in individui asessuati perché ognuno di essi manterrà per sé la propria mutazione favorevole ed inoltre entrerà in competizione con l'altro. Quando uno dei due individui prevarrà, una delle due mutazioni favorevoli sarà persa.

Mendel intuì la diploidia (due alleli di uno stesso gene), la dominanza e la recessività di una memoria chimica (successivamente chiamata gene) che veniva trasmessa dai genitori ai figli.

La 1a legge di Mendel (nella visione delle conoscenze attuali) dice che alla meiosi i due alleli di uno stesso gene segreghino (si separano) per produrre gameti aploidi. I 4 gameti hanno la stessa probabilità di fondersi con un altro gamete dell'altro genitore ed è a caso che uno dei 4 gameti darà luogo alla successiva generazione.

La 2a legge di Mendel dice che alleli di differenti geni segregano indipendentemente. Questo è vero se i due geni sono posti su cromosomi diversi. Alla meiosi i loro alleli si distribuiscono indipendentemente nei singoli gameti. Esempio: un padre riceve da suo padre i cromosomi 1 e 2, alleli paterni (P), e i cromosomi omologhi 1 e 2 da sua madre, alleli materni (M). Alla meiosi, con la distribuzione casuale dei cromosomi, il padre con uguale probabilità

genererà gameti che hanno ambedue i cromosomi 1P e 2P e 1M e 2M o combinazioni di P ed M (1P e 2M; 2P e 1M)(figura 3-6).
Si assume che Mendel abbia formulato la sua 2a legge senza eccezioni perché casualmente analizzò solamente il fenotipo di geni posti su cromosomi diversi (o forse sullo stesso cromosoma ma distanti tra loro, vedere dopo).
Infatti quando due geni sono posti sullo stesso cromosoma la 2a legge di Mendel non è più valida; in particolare quando i geni hanno loci molto vicini non ricombinano mai (sono associati). La frequenza di ricombinazione e quindi della segregazione degli alleli nei gameti aumenta con la distanza fisica tra i due geni e solo se i due geni sono posti vicino ai telomeri (un gene vicino ad un telomero e l'altro all'altro telomero) ricombinano con la stessa frequenza dei geni posti su cromosomi diversi (capitolo 3)
Anche se la deviazione quantitativamente maggiore dalla 2a legge di Mendel è data dai geni geneticamente associati, esistono anche altri meccanismi genetico-molecolari che portano a non rispettare la 2a legge di Mendel. Ad esempio i fenotipi normali e patologici dipendenti da più geni (caratteri poligenici). In questo caso la deviazione dipende dall'osservare un fenotipo che dipende da più geni, perché i singoli geni seguono le leggi di Mendel e le deviazioni date dalla ricombinazione omologa. Altre deviazioni possono provenire dalla comparsa o scomparsa di un allele conseguente a mutazioni spontanee, dalla duplicazione dei geni seguita da mutazione, dalla perdita di eterozigosi nelle cellule della linea germinale, ed inoltre dal mosaicismo e chimerismo e dall'impronta genomica. Alcune di queste eccezioni si verificano per effetto di eventi spontanei pertanto rendono imprevedibili le conseguenti alterazioni del fenotipo (vedere paragrafi successivi: 1-10).

Cause dell'eterogeneità genetica dell'uomo

1. *Distribuzione casuale dei cromosomi omologhi, ricombinazione genetica omologa e riproduzione sessuata*

Il corredo cromosomico aploide di ogni gamete è geneticamente diverso da quello di ogni altro gamete prodotto dallo stesso individuo perché durante la meiosi si ha la distribuzione casuale dei cromosomi omologhi e la ricombinazione genetica omologa (anche detta ricombinazione omologa, ricombinazione generale o crossing-over).
La riproduzione sessuata, con la fecondazione (unione del gamete maschile con quello femminile e formazione dello zigote) determina la combinazione dei corredi cromosomici aploidi del padre e della madre che nella specie umana si assume siano diversi tra loro per circa il 10% degli alleli. Quindi con due meccanismi (molecolari e cellulari) diversi, prima si ricombina il corredo genetico diploide di ciascun individuo genitore (distribuzione casuale dei cromosomi e ricombinazione omologa alla meiosi), poi il corredo aploide dei due individui diversi (genitori) si combina (fecondazione) per formare un nuovo individuo (figlia/o).
In conseguenza di ciò i figli di una stessa coppia risultano geneticamente diversi tra loro, diversi dai genitori, da tutti gli altri ascendenti e discendenti e dagli

altri individui della stessa specie. Quindi il mantenimento dell'eterogeneità genetica di tutti gli individui di una data specie dipende principalmente dai meccanismi sopra indicati.
Nelle gonadi durante la meiosi, i singoli cromosomi omologhi (di origine paterna e materna) si distribuiscono nelle cellule figlie casualmente ed indipendentemente l'uno dall'altro. Nella specie umana risultano 2^{23} = 8.388.608 possibili differenti combinazioni di cromosomi nei gameti prodotti da ciascun genitore. Le diverse combinazioni dei cromosomi alla formazione dello zigote sono 2^{46}. Altre variazioni sono date dalla ricombinazione omologa meiotica (crossing-over). Si stima che nella specie umana, durante ogni meiosi si abbia almeno una ricombinazione omologa per ogni cromosoma (nella donna la ricombinazione genica è più frequente che nell'uomo) e che le differenze di alleli tra padre e madre interessino circa il 10% dei geni (vedere i paragrafi 2 e 3). Risulta così che il numero dei possibili zigoti geneticamente diversi di ogni singola coppia è maggiore di $6x10^{43}$. Questo numero è superiore a quello di tutti gli esseri umani vissuti fino ad oggi e testimonia l'unicità genetica di ogni individuo. Dato il basso numero di figli che una coppia può mettere al mondo, è (umanamente!) impossibile che essa generi due gemelli genetici con due zigoti diversi e a maggior ragione è impossibile trovare gemelli generati da coppie diverse di genitori. Anche i romantici sanno che l'essere amato è unico al mondo (non se ne può trovare un altro uguale, salvo fenocopie). Si calcola che individui diversi abbiano alleli diversi solo nel 10% degli stessi loci genici.
La base culturale del razzismo umano, cioè lo scarso rispetto delle differenze di fenotipo esistenti tra individui diversi, sembra essere generato dalla grande considerazione data alle poche differenze fenotipiche tra le varie razze piuttosto che all'unicità del genoma della popolazione umana, alla natura dei meccanismi che provocano l'eterogeneità genetica nell'uomo ed alla omologia ed alta similarità genetica degli individui appartenenti a razze diverse. Tutto il problema sta in piccole differenze. Se poi si pensa che la similarità delle sequenze del DNA umano con quelle dello scimpanzé è circa il 98-100%, è un po' assurdo che magari ci sentiamo più simili allo scimpanzé che non al vicino di casa di un'altra "razza". Altra rivalità, le differenze tra donna e uomo, dove l'uomo differisce dalla donna soprattutto per l'azione del gene SRY. Un gene in più, grandi differenze di fenotipo: da una donna simile a Monica Bellucci e per attività del gene SRY può venire fuori un uomo simile a Francesco Totti.

2. Polimorfismo genetico.

Il polimorfismo genetico si verifica quando in una popolazione di individui sani lo stesso gene (steso locus) può esistere in due o più forme tutte codificanti la stessa proteina. Le forme di uno stesso gene sono dette alleli e per definizione hanno tutti lo stesso locus: la localizzazione subcromosomica fisica. Il polimorfismo può essere individuato osservando la diversità di uno stesso carattere fenotipico o la diversità della sequenze di uno stesso locus. I genetisti tradizionali valutavano la presenza di alleli di uno stesso gene sulla base di variazioni del fenotipo (es. gruppi del sangue, mobilità elettroforetica di alcuni

enzimi) e statisticamente stabilivano l'esistenza di polimorfismo in un dato locus quando un allele aveva la propria frequenza superiore all'1%. Per valori inferiori si assumeva che l'allele fosse continuamente generato da mutazioni ricorrenti e non trasmesso geneticamente da una generazione all'altra.
I genetisti molecolari che analizzano le sequenze del DNA genomico nucleare stabiliscono la presenza di polimorfismo in un dato locus se una sequenza, codificante o non codificante, variante anche per una singola base (SNP), è presente nel DNA di un singolo individuo di una intera popolazione.
Le nuove tecnologie del DNA hanno permesso di individuare molti loci polimorfici di sequenze non codificanti: RFLP, VNTR, SNP (capitolo 3).
L'alta densità nel genoma dei loci SNP (uno ogni 1000b) suggerisce che tutti i geni umani siano polimorfici.
L'esatta valutazione del polimorfismo umano si avrà quando sarà nota la sequenza degli alleli di tutti i geni presenti nella popolazione umana.
Le sequenze EST che sono l'aspetto molecolare del polimorfismo dei geni umani vengono conservate in una specifica banca dati.
L'analisi del polimorfismo dei geni umani osservato attraverso il loro fenotipo indicava che solo il 30% di essi era polimorfico. Il contrasto tra i dati biochimici (fenotipo proteico) e quelli molecolari del DNA (genotipo) è dato dal fatto che con l'analisi del DNA si valutano anche gli alleli che pur avendo tra loro una sequenza nucleotidica diversa, per la degenerazione del codice genetico, codificano la stessa sequenza aminoacidica (la proteina ha la stessa mobilità elettroforetica). Altri alleli possono produrre una proteina modificata di uno o due aminoacidi che tuttavia non modificano né la carica né la attività molecolare della proteina perché gli aminoacidi sono simili (sostituzione conservativa, es. Asp con Glu) specialmente se sono posti sulla superficie della proteina. Con questa sostituzione le due forme di proteina risultano identiche per mobilità elettroforetica ed attività molecolare della proteina.

Polimorfismo dei geni e proprietà minori o subdole delle proteine.
Alcuni alleli di uno stesso gene possono codificare proteine con variazioni di sequenza di pochi aminoacidi (non necessariamente vicini), ma in genere codificano proteine con variazioni di un singolo aminoacido che a livello genico risulta dal polimorfismo di una singola base (SNP). Queste proteine hanno una normale attività molecolare ma valori maggiori o minori di proprietà che contribuiscono alla normale attività molecolare oppure hanno proprietà accessorie nuove. Le prime proprietà includono: la stabilità della proteina (vita della proteina nella cellula), valore della Vmax per gli enzimi, il grado di sensibilità alle molecole segnale (effettori, ormoni; le seconde proprietà includono:,grado di sensibilità a composti esogeni (farmaci, anestetici, allergeni, conservanti degli alimenti, inquinanti). Queste proprietà sono dette minori (modificano poco la attività molecolare e la funzione della proteina), o subdole quando si manifestano impreviste in particolari condizioni ambientali o di alimentazione/somministrazione con effetti negativi inaspettati.

Le proprietà minori non sono solo negative, come la sensibilità ai farmaci che permette di curare/guarire stati patologici, o l'insensibilità ad allergeni mentre sono proprietà minori negative, ad esempio, la sensibilità agli allergeni, agli inquinanti, la sensibilità negativa ai farmaci (effetti collaterali).
La sensibilità a composti esogeni è in genere una proprietà subdola perché si manifesta quando un individuo è costretto a fare uso di farmaci , anestetici o introduce involontariamente nel suo organismo allergeni, conservanti o inquinanti perché cambia alimentazione o si sposta da un ambiente ad un altro. Inoltre in relazione alla costituzione genetica di alleli di uno stesso gene, individui diversi manifestano sensibilità diverse verso la stessa molecola esogena (figura D-7).
Le carenze genetiche di attività catalitica causate da una minore concentrazione di enzima e/o del suo numero di turnover possono essere compensate a livello del fenotipo mediante meccanismi di regolazione degli enzimi, tuttavia con perdite di potenzialità metabolica (figuraD-6).
Si ritiene che alcune proprietà minori siano responsabili della variazione quantitativa di alcune caratteristiche individuali multifattoriali (dipendenti da più geni, da fattori ambientali e dall'alimentazione) come l'altezza corporea, la pressione del sangue, la concentrazione delle proteine e dei metaboliti nelle cellule e nei fluidi biologici. Il carattere subdolo delle proprietà minori di alcune proteine che concorrono a formare un carattere multifattoriale normale, sta nel fatto che esso può essere responsabile di caratteri patologici multigenici in relazione ad alcune, ma non tutte, le combinazioni di forme alleliche delle stesse proteine (capitolo 4 e appendice E).
Gli alleli responsabili di patologie genetiche fanno anch'essi parte del polimorfismo di un gene, tuttavia data l'importanza medica e sociale sono classificati separatamente e se un gene ha più alleli mutati responsabili di una patologia si parla di poliallelismo patologico (appendice E). Quindi un gene in relazione agli effetti sul fenotipo può essere dotato di più forme alleliche: alleli che codificano proteine normali e proteine dotate di proprietà subdole (polimorfismo), e più alleli che codificano la proteina inattiva (poliallelismo patologico).Appare incredibile come con pochi dati a disposizione nel 1902 Sir Archibald E. Garrod abbia predetto che la variazione patologica fosse il caso estremo della variazione normale.
Un tipo particolare di proprietà subdola si riscontra in alleli di geni poliallelici con uno allele patologico recessivo. Uno di questi alleli sia in omozigosi che in eterozigosi con l'allele normale non causa stati patologici. Mentre lo stesso allele in condizioni di eterozigosi con l'allele patologico recessivo, determina in un individuo lo stato patologico (un esempio nell'uomo è l'emoglobina HbC). Questo perché la proteina del primo allele (subdolamente non patologica) ha delle proprietà strutturali (non presenti nelle proteine normali) che le permettono di collaborare con la proteina dell'allele patologico a determinare la patologia, ma le stesse proprietà strutturali non sono sufficienti a causare la patologia anche quando si ha omozigosi dell'allele HbC codificante l'emoglobina subdolamente non patologica. Questo tipo di proprietà subdola è utile per

capire perché alcuni autori parlano di carattere genetico (trait) dominante, recessivo o patologico ed evitano di indicare il relativo allele come dominante, recessivo o patologico. Infatti in alcuni casi definire un allele come patologico può ingannare perché come accade per l'emoglobina HbC, l'allele è patologico o meno in dipendenza dell'altro allele omologo presente nello stesso individuo. In altri casi, il carattere patologico è il risultato delle interazioni del prodotto di un allele con quello di altri geni e/o con fattori ambientali come accade nella dipendenza dall'età della penetranza delle malattie monogeniche dominanti. In tutti questi casi l'allele è sempre lo stesso, ma in relazione alle proteine od altri componenti cellulari o molecole esogene che il suo prodotto (proteina) incontra nelle cellule, manifesta attività diverse (attività normale o patologica, dominante o recessiva). Da ciò l'attribuzione di queste attività al fenotipo (proteina) piuttosto che al genotipo (allele). Tuttavia nella pratica per semplificare si usa dire allele/gene dominante, recessivo o patologico sottintendendo che esprime rispettivamente un carattere dominante, recessivo o patologico sempre, oppure solo in particolari condizioni che devono essere descritte.

3*. Dominanza, recessività e codominanza delle caratteristiche genetiche*.
Alleli normali diversi di uno stesso gene codificano proteine che, espresse nella cellula, svolgono gradi diversi della stessa funzione e possono influenzare l'una l’attività molecolare dell'altra.
Dominante è riferito ad alleli che manifestano il loro fenotipo (il loro carattere genetico, trait) anche in condizioni di eterozigosi, cioè in presenza di un diverso allele omologo (figura D-1). Recessivo è l'allele che si manifesta solo in condizioni di omozigosi (alleli identici nello stesso locus). L'allele dominante può differire dall'allele recessivo anche per una singola base. Codominante è la condizione di eterozigosi in cui si possono osservare le caratteristiche di ambedue gli alleli (Es. il gruppo del sangue AB risulta dalla codominanza degli alleli A e B). Un caso particolare di dominanza, detta dominante negativa, si ha quando il prodotto genico di un allele prevale perché inibisce l'azione del prodotto genico normale dell'altro allele espresso nella stessa cellula.
Gli alleli recessivi che codificano enzimi con carente attività catalitica per le normali funzioni metaboliche (causate da una minore concentrazione di enzima e/o del suo numero di turnover), in eterozigosi non causano alterazionipatologiche quando la carenza dell'attività catalitica é compensata a livello del fenotipo mediante meccanismi di regolazione degli enzimi, tuttavia si verificano perdite di potenzialità metabolica (figura D-6).

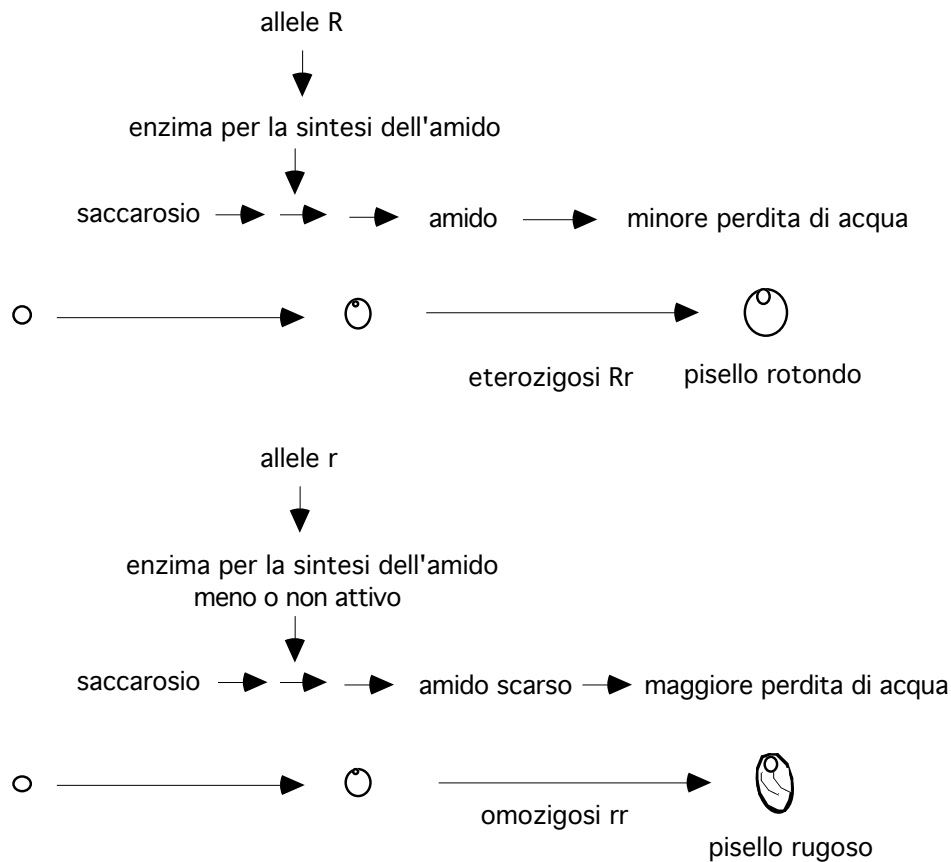



Figura D-1. Basi genetico-molecolari della formazione dei piselli lisci e rugosi.
L'allele R è dominante e codifica un enzima della via metabolica di sintesi dell'amido da saccarosio; l'allele r è recessivo e codifica un enzima meno o non attivo. La trasformazione del saccarosio in amido fa assorbire un data quantità di acqua al pisello rotondo, alla maturità questa acqua viene quasi tutta mantenuta pertanto il pisello mantiene la sua forma rotonda. Nel pisello rugoso è presente una maggiore quantità di saccarosio (data la ridotta trasformazione dello zucchero in amido) e questo fa assorbire molta acqua al pisello. Successivamente alla maturità l'acqua viene in gran parte persa rendendo il pisello rugoso(da Genetica. Un approccio molecolare, Brown T. A., Piccin, 2002, ridisegnato e modificato).

----

4. *Caratteri poligenici e multifattoriali*.

Certi caratteri del fenotipo umano dipendono dall'interazione del prodotto (proteina) di più geni posti su loci indipendenti. Si assume che il prodotto di ognuno di questi geni dia un contributo al carattere poligenico ma che esso si sommi con quello degli altri geni e dalla loro somma risulti il carattere poligenico. Il carattere che dipende dall'interazione di più geni è detto multifattoriale quando la sua manifestazione dipende anche da uno o più fattori ambientali (talvolta i due termini poligenico e multifattoriale sono usati come sinonimi). Si ipotizza che gli stessi geni responsabili del carattere poligenico rendano l'individuo più sensibile ai fattori ambientali, perché dipendendo da più proteine si ha una maggiore possibilità che una di esse possa associare molecole esogene (ambientali e farmaci).

I caratteri multifattoriali possono essere continui, cioè avere una gradualità continua di valori, o discontinui, cioè essere presenti o assenti. Caratteri umani

multifattoriali continui sono: l'altezza, il peso, l'intelligenza, la pressione arteriosa, il colore della pelle, la dimensione dei globuli rossi. I caratteri discontinui umani possono essere anche malformazioni congenite (es. il labbro leporino, le malattie cardiache congenite, la stenosi pilorica) o patologie che colpiscono gli adulti (es. diabete mellito, epilessia, schizofrenia).
I caratteri multifattoriali risultano da combinazioni di alleli di più geni, dove il prodotto dei singoli alleli può avere un peso diverso in individui diversi (portatori di combinazioni diverse di alleli dello stesso gruppo di geni), questa è una eterogeneità diversa da quella dei caratteri monogenici che esprimono caratteri discontinui (presenti o del tutto assenti).

5. *Mutazioni spontanee*
La frequenza delle mutazioni spontanee è circa 10^{-6}/gene/ciclo cellulare *in vitro*, anche in un ambiente privo di mutageni. Nell'uomo si è calcolato che avvengano 5×10^{-2} mutazioni per gamete (genoma aploide), cioè la probabilità di avere una mutazione è 1/20 e per lo zigote è 1/10. Durante la vita nel corpo umano si verificano complessivamente circa 10^{16} divisioni cellulari per cui ogni singolo gene (non della stessa cellula) ha teoricamente circa 10^{10} occasioni indipendenti di subire una mutazione ed è stato calcolato che debbano occorrere almeno 3-7 mutazioni per trasformare una singola cellula in tumore.
Nell'uomo circa il 23% delle mutazioni sono silenti (non alterano il fenotipo). In questo paragrafo come causa di eterogeneità genetica sono considerate solo le mutazioni silenti e quelle che causano effetti positivi all'organismo. Mentre le mutazioni che causano effetti negativi all'organismo sono responsabili dell'eterogeneità patologica del genoma umano. Tuttavia, le cause che instaurano ogni tipo di mutazione (silenti, patologiche, ecc.) sono le stesse, la differenza è nell'alterazione che la mutazione determina nella sintesi e/o nella struttura del RNA (tRNA, rRNA) o della proteina codificata e quindi nella concentrazione e/o nell'attività molecolare e funzione fisiologica di queste macromolecole. L'alterazione causata dalla mutazione può risultare in una attività molecolare inalterata, persa, eccessiva, carente o nuova ai fini della fisiologia della cellula e dell'organismo.

6. *Potenziamento del genoma (aggiunta di nuovi geni al genoma).*
a). La duplicazione di un gene, seguita da mutazione di un allele dello stesso gene, è un meccanismo genetico importante perché, dato il lungo cammino evolutivo percorso dall'uomo, si ritiene che le nuove mutazioni, anche se conferiscono nuove proprietà positive alla proteina espressa dal gene mutato, in genere ne riducono altre già presenti nella proteina normale. Ad esempio, quando una proteina mutata è meno stabile, ha una vita più breve e la cellula deve spendere più energia per mantenere la proteina mutata nella stessa concentrazione della proteina normale. Tuttavia la regolazione della concentrazione e quindi della attività molecolare della proteina mutata risulta più pronta di quella della proteina normale. Infatti quando l'espressione della

proteina viene inibita, la proteina mutata scomparirà dalla cellula più rapidamente della proteina normale.
La proteina normale, la cui concentrazione cellulare è regolata più lentamente, garantisce gran parte dell'attività molecolare di base (es. catalisi di una reazione metabolica). La proteina mutata, la cui concentrazione è regolata più prontamente, permette una migliore regolazione in risposta alle esigenze della fisiologia cellulare. Se ipotizziamo che il 50% dell'attività catalitica di base sia dato dall'enzima normale ed il rimanente 50% da quello mutato, quando l'espressione genica viene inibita per ambedue i geni (es. hanno lo stesso promotore), per il 50% dell'attività molecolare ci sarà una caduta più rapida data la labilità della proteina mutata. Anche la sintesi della proteina è più rapida data la presenza di due geni.
b). L'inserimento casuale nella cellula di plasmidi o l'integrazione nel DNA cellulare di geni plasmidici o virali possono conferire alla cellula nuove funzioni fisiologiche o patologiche.
c). Amplificazione genica. L'amplificazione genica è l'incremento del numero di copie di un gene che incrementa la potenzialità funzionale della cellula. Il trattamento di cellule in cultura con methotrexate, inibitore dell'enzima deidrofolato-riduttasi causa la morte di molte cellule, sopravvivono solo quelle che hanno incrementato il numero dei geni dell'enzima deidrofolato-riduttasi. L'enzima supera l'inibizione causata da una data concentrazione di methotrexate perché la sua concentrazione cellulare è aumentata (effetto Mitridate). L'amplificazione interessa anche geni vicini a quello dell'enzima deidrofolato-riduttasi.
Tuttavia l'amplificazione genica può creare problemi in alcuni tipi di terapia, perché può essere indotta da farmaci per geni che codificano enzimi deputati alla detossificazione di composti esogeni. L'incremento di questi enzimi provoca una più rapida inattivazione/eliminazione del farmaco usato instaurando la resistenza al farmaco stesso (il farmaco è meno attivo sul paziente perché è subito modificato/eliminato).

7. *Ricombinazione sito specifica e trasposizione di DNA.*
La ricombinazione sito-specifica è lo scambio di segmenti di DNA a livello di specifiche sequenze con formazione di nuovi geni (es. la formazione dei geni delle immunoglobuline). La trasposizione interessa piccoli segmenti di DNA (trasposoni) che hanno la capacità di spostarsi in regioni diverse di uno stesso cromosoma o di cromosomi diversi. Anche questo tipo di ricombinazione può portare alla formazione di nuovi geni.

8. *Perdita di eterozigosi* (LOH = Loss of Hetherozygosity).
E' una modifica del genoma che contrariamente a quelle sopra descritte, porta ad una riduzione di eterogeneità genetica nella cellule germinali e somatiche. Mediante più tipi di meccanismo (figure D-2 e D-3), un cromosoma o un gene o parti di essi sono resi uguali al rispettivo omologo. Al termine di uno dei meccanismi della LOH, la cellula risulta avere un cromosoma o un gene (o parti

di esso) identici (ambedue di origine materna o paterna) mentre prima della LOH erano diversi (eterozigoti).
La frequenza della LOH ($10^{-6}\div10^{-3}$/gene/ciclo cellulare) in genere è più alta di quella delle mutazioni spontanee (circa 10^{-6}/gene/ciclo cellulare).
I meccanismi della LOH sono:
a). Perdita di un cromosoma che può essere seguita dalla duplicazione del cromosoma residuo (figura D-2).
b). Ricombinazione omologa mitotica. Due cromosomi omologhi si scambiano reciprocamente una parte (figura D-3 a). Nelle cellule somatiche la ricombinazione generale risulta in una perdita di eterozigosi perché, dopo la divisione cellulare (mitosi), le cellule figlie rimangono diploidi e posseggono i due cromosomi omologhi (materno e paterno) dei quali quelli che hanno subito la ricombinazione hanno una regione identica (figura D-3a).
Nella meiosi la ricombinazione incrementa l'eterogeneità dei gameti perché dopo la ricombinazione i cromosomi omologhi ricombinati si separano finendo in cellule (gameti) diverse che hanno corredi aploidi, cioè singole copie di cromosomi (ricombinati o no). In conseguenza di ciò l'eterogeneità dei cromosomi risulta incrementata e di conseguenza anche quella dei gameti. Osserviamo la figura D-3a, assumiamo che le cellule H e K siano gametociti e che vadano incontro a meiosi. Da esse si formeranno 4 gameti diversi proprio in virtù del crossing-over, mentre senza ricombinazione i due cromosomi M ed i due cromosomi P sarebbero rimasti uguali.
c). Conversione genica.
Si verifica durante ricombinazioni meiotiche e mitotiche in regioni dove le sequenze materne e paterne sono leggermente diverse per cui nelle giunzioni eteroduplici l'appaiamento dei filamenti (materno e paterno) non è completo. Successivamente i meccanismi di riparazione correggono la basi non appaiate ricopiando l'uno (materno) o l'altro (paterno) (figure D-3b e D-3c). A seconda di quale filamento è riparato, al termine della meiosi si possono avere tre gameti (invece di due) con un gene di tipo materno ed il quarto di tipo paterno (o viceversa tre di tipo paterno ed uno di tipo materno). Nella mitosi, una cellula figlia diviene omozigote per il gene che è stato convertito e l'altra cellula rimane eterozigote come lo era la cellula madre.
Nella meiosi la ricombinazione omologa è più frequente che nella mitosi, ne consegue che anche la conversione genica è più frequente che nella meiosi che nella mitosi. Ciò appare in relazione alle caratteristiche fisiche e di funzione dei due meccanismi: nella meiosi c'è una fase specifica per l'appaiamento dei cromosomi omologhi (diplotene, che è assente nella mitosi) in cui avviene la ricombinazione, la meiosi ha tra le sue funzioni quella di diversificare il genoma, mentre la mitosi (normale ciclo cellulare) quella di duplicarlo, di dividerlo e di trasmetterlo alle cellule figlie così come lo ha ricevuto.
d). Mutazione inversa. E' la mutazione che rigenera l'allele normale o la mutazione puntiforme (o delezione) che sopprime l'espressione di un allele (figura D-2). La rigenerazione dell'allele normale può avvenire anche per LOH.

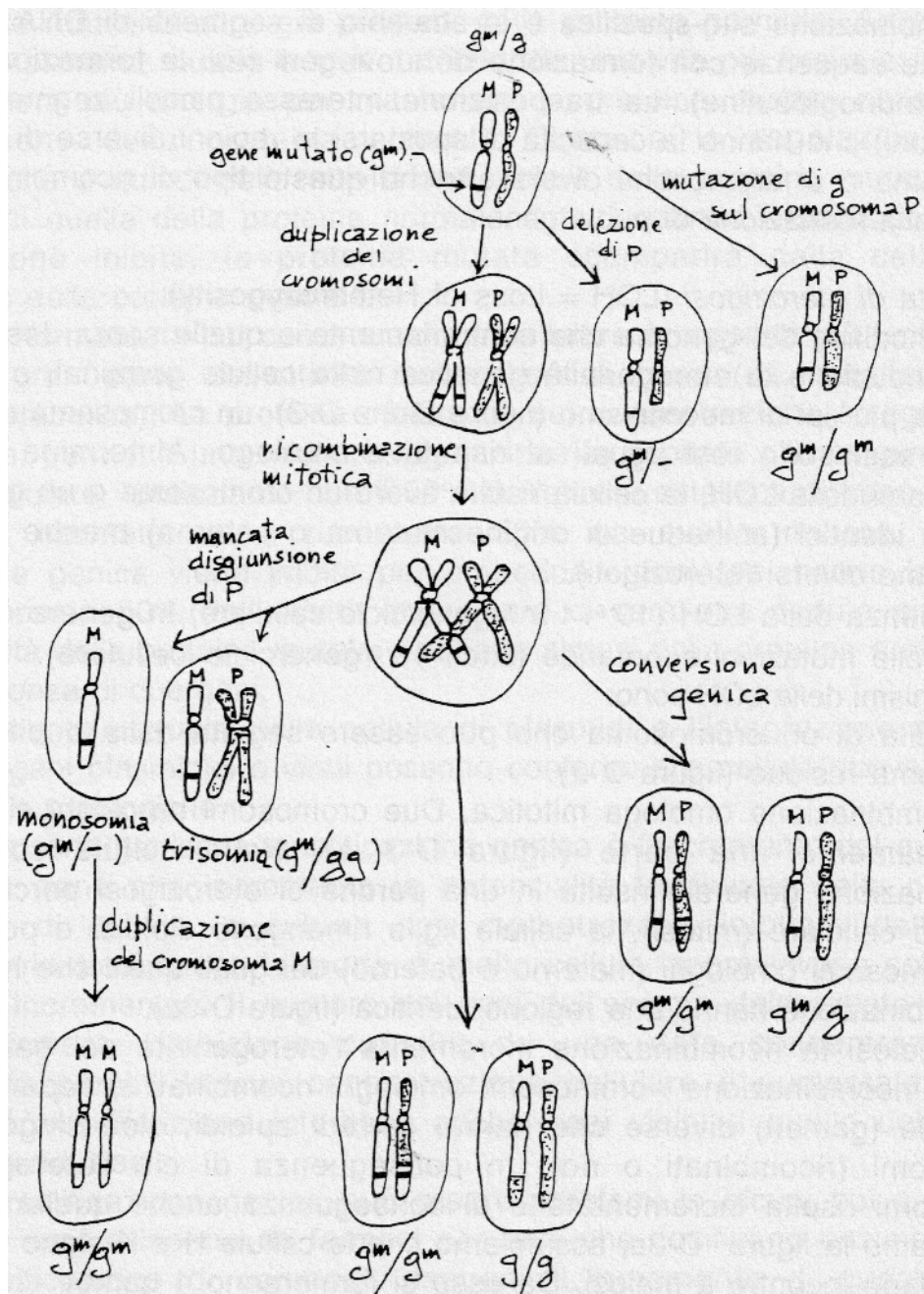



Figura D-2. Perdita di eterozigosi (LOH).

In figura sono schematicamente indicati i meccanismi tramite i quali si può instaurare la LOH con il conseguente instaurarsi della omozigosi (g^m/g^m) di un gene mutato (g^m) portatore di una patologia. In due casi (delezione e monosomia) la LOH si instaura per perdita del gene g (normale) e la patologia si manifesta perché la carenza del prodotto genico del gene g^m non è più compensata dal prodotto del gene normale g^m. M, materno e P, paterno sono cromosomi omologhi. I meccanismi della LOH operano anche per alleli diversi non patologici di uno stesso gene.

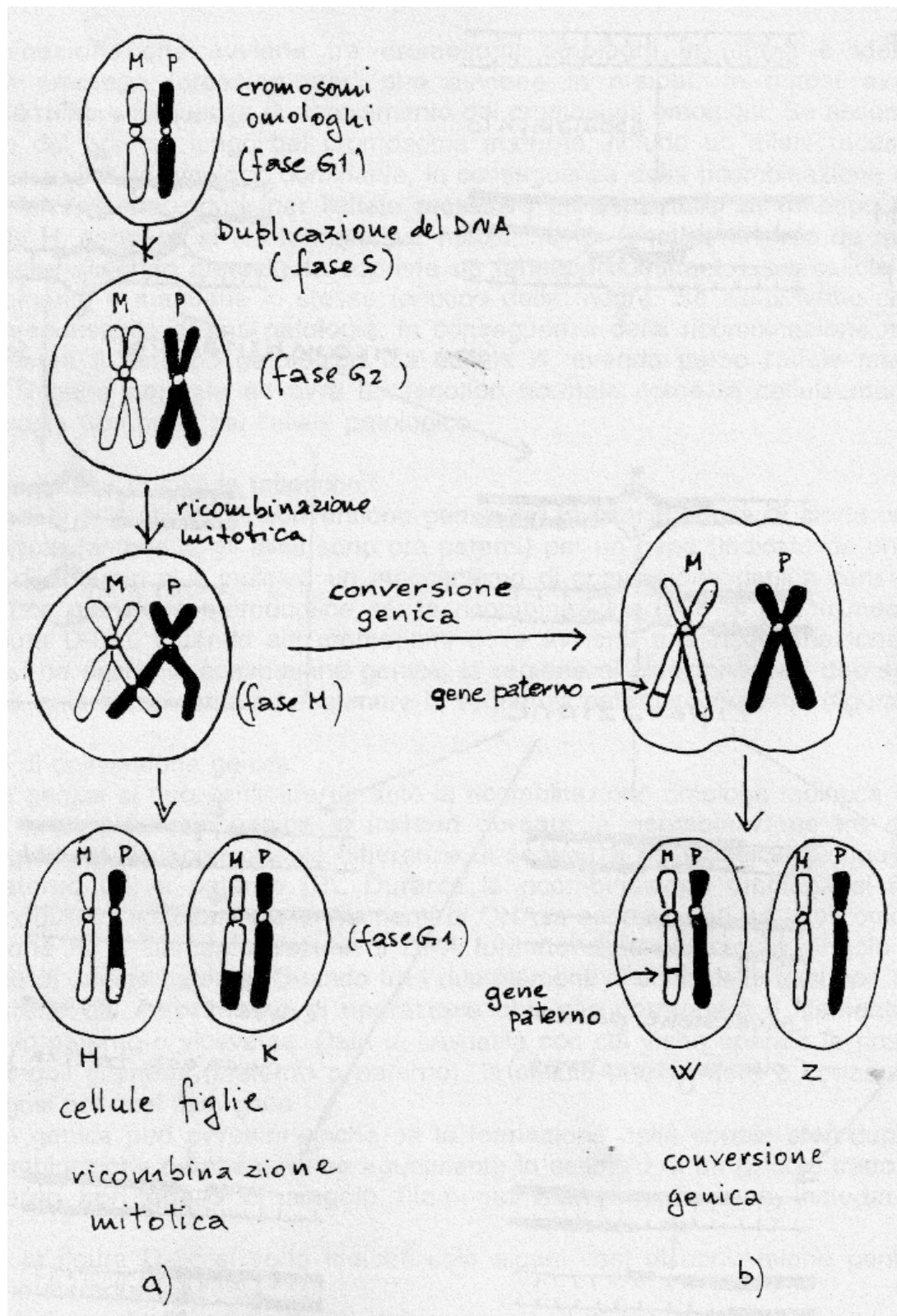



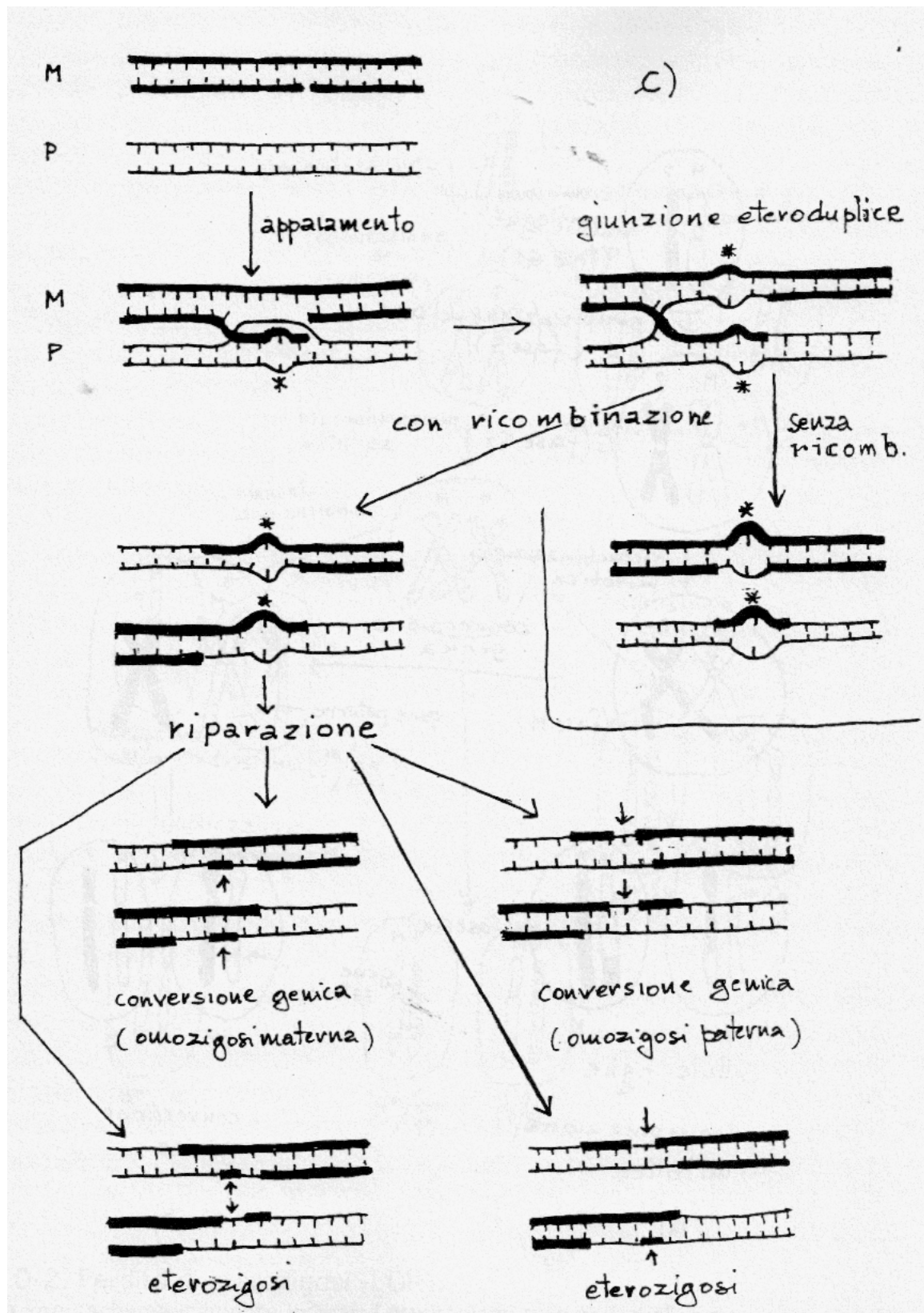
Figura D-3 a. Ricombinazione mitotica; b. Conversione genica.

FiguraD-3. c. Meccanismi molecolari di conversione genica.

Figura D-3. Ricombinazione mitotica e conversione genica.

a). La ricombinazione che avviene tra cromosomi omologhi in mitosi è identica alla ricombinazione omologa (crossing-over) che avviene in meiosi. In mitosi avviene più raramente per la minore frequenza di appaiamento dei cromosomi omologhi. Se assumiamo che la parte distale del braccio lungo del cromosoma materno includa un allele recessivo ed il cromosoma paterno l'allele omologo dominante, in conseguenza della ricombinazione mitotica la cellula figlia H diverrà omozigote per l'allele recessivo ed esprimerà un fenotipo recessivo. Quindi la cellula H, sebbene si sia formata per mitosi, ha un fenotipo diverso da quello della cellula madre (che essendo eterozigote esprime un fenotipo dominante). La cellula figlia K è omozigote dominante e mantiene lo stesso fenotipo della madre. Se assumiamo che il gene recessivo sia responsabile di una patologia, in conseguenza della ricombinazione mitotica, la cellula H esprimerà il fenotipo patologico. La cellula K, avendo perso l'allele mutato, sarà omozigote per il gene normale ed avrà un fenotipo normale come la cellula madre, ma a differenza di questa non porta più l'allele patologico.

b). Conversione genica in cellule mitotiche.
Durante la mitosi si è verificata la conversione genica ed in conseguenza di ciò la cellula W è divenuta omozigote (ambedue gli alleli sono ora paterni) per un gene (indicato da una freccia). Per semplicità di disegno si è indicato un meccanismo di conversione genica conseguente la formazione di una giunzione eteroduplice senza ricombinazione (uno di questi meccanismi è descritto in figura D-3c). Quando sui cromosomi dove avviene una ricombinazione omologa (figura D-3a) si ha anche la conversione genica, la regione di giunzione tra i due segmenti di cromosoma può essere convertita ed assumere la sequenza paterna o materna (figura D-3c).

c). Meccanismi di conversione genica.
La conversione genica si può verificare durante la ricombinazione omologa meiotica e mitotica. Si ritiene che la conversione genica si instauri durante la ricombinazione tra cromosomi omologhi in regioni di DNA con piccole differenze di sequenza nucleotidica (alcune basi) tra i cromosomi materno (M) e paterno (P). Durante la ricombinazione omologa si forma una giunzione eteroduplice (appaiamento di filamenti di DNAss appartenenti ad alleli omologhi). La zona di giunzione ha il filamento doppio di DNA formato da un filamento singolo di origine materna e l'altro di origine paterna. Quando tra i due filamenti ci sono delle basi non accoppiate esse sono corrette dal meccanismo di riparazione che può correggere il filamento materno ricopiando quello paterno o viceversa. Data la casualità con cui viene operata la riparazione di uno dei due singoli filamenti (materno o paterno), la cellula può perdere o conservare il suo stato di eterozigosi per quel dato gene.
La conversione genica può avvenire anche se la formazione della coppia eteroduplice non è seguita da ricombinazione, perché avviene egualmente lo scambio di un piccolo tratto di singolo filamento paterno con quello di singolo filamento materno che può includere le basi disaccoppiate.
Per semplicità in figura D-3c si sono indicati solo alcuni casi di conversione genica di una stessa giunzione eteroduplice.
Oltre a quello indicato in figura, esistono altri meccanismi di conversione genica, anche essi chiedono la formazione di giunzioni eteroduplici tra filamenti di DNA con piccole differenze di sequenza e successive reazioni di riparazione.

---

* indica una sola coppia di basi disaccoppiate tra il filamento di DNAss M e quello P (in realtà possono essere più di una, non necessariamente vicine); ↑↓, le piccole frecce verticali indicano la base riparata.

---

d). Mutazione inversa. E' la mutazione che rigenera l'allele normale o la mutazione puntiforme (o delezione) che sopprime l'espressione di un allele (figura D-2). La rigenerazione dell'allele normale può avvenire anche per LOH.
Con i meccanismi indicati in a), c) e d) nelle cellule somatiche si ha la stessa probabilità teorica di mantenere solo l'allele paterno o solo quello materno. Con il meccanismo descritto in b) dopo la divisione cellulare, una cellula figlia sarà omozigote per un gene paterno e l'altra per il gene materno. Quindi con i meccanismi indicati in a), c) e d), un allele mutato patologico ha il 50% di possibilità di essere eliminato e 50% di essere reso omozigote. Mentre con il meccanismo descritto in b) è eliminato in una cellula e reso omozigote nell'altra.
Risulta così che la LOH può instaurare una malattia genetica quando rende omozigote un gene responsabile di una patologia recessiva o rendere più grave una malattia genetica rendendo omozigote un gene responsabile di una patologia dominante. Tuttavia per poter instaurare una patologia la LOH deve avvenire in un giovane embrione al fine che molte cellule dell'adulto (es. tutte quelle di un organo) abbiano il DNA modificato dalla LOH. La LOH può avere conseguenze drammatiche anche quando interessa il DNA di una singola cellula somatica se rende omozigote un antioncogene mutato, perché può instaurare la cancerogenesi. Le regioni cromosomiche dove avviene più frequentemente la LOH sono studiate perché si assume che possano indurre la cancerogenesi recessiva rendendo omozigote l'allele di un antioncogene mutato.
Il retinoblastoma è un tumore pediatrico provocato dall'antioncogene Rb mutato in condizioni di omozigosi (cancerogenesi recessiva). Quando un allele Rb è ereditato mutato dai genitori, nel 95% dei neonati a causa della perdita della eterozigosi (LOH) si ha la formazione di uno-tre retinoblastomi in ciascun occhio. La retina in formazione ha circa $4x10^6$ cellule ed in essa avviene un evento di LOH ogni 10^6 cellule (circa). LOH ha una frequenza di circa 10^{-4} eventi/gene/replicazione cellulare. Risulta così che il retinoblastoma familiare abbia la frequenza di 1/10.000 neonati mentre quello sporadico, che è causato da due mutazioni spontanee, una per ciascun allele (10^{-6} mutazioni/gene/replicazione cellula), ha una frequenza di 1/30.000 neonati.
9. *Mosaicismo e chimerismo.*
Un individuo mosaico è un individuo originato da un unico zigote avente due o più linee cellulari geneticamente diverse per uno o più geni o interi cromosomi.
Un individuo chimera è un individuo originato da più di uno zigote.
Il mosaicismo è un fenomeno comune e può interessare la linea germinale come quella somatica. Esso causa difficoltà nell'interpretare gli alberi genealogici, mentre il chimerismo è molto raro.
L'individuo mosaico risulta da:
a) mutazioni post-zigotiche, mutazioni che avvengono in una cellula di un embrione allo stadio di due o più cellule, cioè in fasi precoci dello sviluppo embrionale di maschi e femmine. Se la mutazione avviene tardivamente nell'individuo le cellule mutate interesseranno solo una parte del corpo od un organo. Un caso particolare di mosaicismo somatico sono i tumori.

b) Inattivazione del cromosoma X nelle femmine (anche detto Lyon effect).
Precocemente durante lo sviluppo embrionale di organismi femminili di mammiferi, donna inclusa, uno dei due cromosomi X delle cellule somatiche è reso geneticamente inattivo (esclusa una piccola zona nella parte distale del braccio piccolo del cromosoma).
L'inattivazione è causata dal gene XIST che codifica RNA di grandi dimensioni il quale si associa al DNA del cromosoma X inibendo tutti o quasi tutti i suoi geni. Con un meccanismo ancora ignoto ed apparentemente a caso, è attivo il gene XIST di un solo cromosoma X ed il suo RNA non migra sull'altro cromosoma X.
Il cromosoma inattivato rimane condensato durante gran parte dell'interfase ed è indicato come corpo di Barr. Nella donna, l'inattivazione avviene circa 16 giorni dopo la fecondazione quando l'embrione è costituito da circa 5.000 cellule e le cellule staminali emopoietiche sono solo 3-5. In ogni singola cellula i due cromosomi X hanno la stessa probabilità di essere inattivati e quando in una cellula viene inattivato uno dei due cromosomi X (es. paterno) esso rimane inattivo in tutte le cellule che originano da quella cellula. In una cellula (vicina alla prima) può essere inattivato il cromosoma X di origine materna, di conseguenza il corpo della donna risulta essere un <u>mosaico</u> di cellule delle quali una parte ha geneticamente attivo il cromosoma X paterno e l'altra parte di cellule il cromosoma X materno. Se uno dei due cromosomi X (es. materno) include una mutazione patologica, solo la parte del corpo che ha le cellule con il cromosoma X materno attivo mostrerà tale malattia. Data la casualità del processo di inattivazione, la proporzione e la disposizione nell'organismo delle cellule che portano il gene mutato varia da donna a donna, anche tra gemelle geneticamente identiche (monozigoti). Può anche esserci una selezione positiva o negativa dei cloni cellulari (contenenti il cromosoma X con il gene mutato) influenzata dal prodotto del gene mutato stesso. Se il gene patogeno causa la morte cellulare, le cellule che sopravvivono sono solo quelle il cui cromosoma X contiene il gene normale. Di conseguenza l'organismo finisce per avere solo cellule con il gene normale. In altri casi, le cellule con il cromosoma X che contiene il gene patologico, sopravvivono perché sono in comunicazione mediante aperture organizzate della membrana plasmatica (gap junctions) con le cellule normali a loro vicine. Da esse ricevono metaboliti di cui sono carenti per incapacità di sintesi a causa della mutazione di un dato enzima. Attraverso questi due meccanismi di compensazione gli organismi femminili possono avere una vita normale.
La distrofia muscolare di Ducenne è una forma grave recessiva di distrofia muscolare. Il gene è sul cromosoma X, per cui la patologia risulta dominante nei maschi. Le madri dei ragazzi affetti da questo tipo di distrofia possono avere manifestazioni lievi della malattia (es. polpacci ipertrofici, debolezza degli arti). Si ritiene che nelle fibre muscolari alterate di queste donne sia attivo il cromosoma X mutato nel gene della predisposizione alla distrofia muscolare di Ducenne. Un esempio dell'espressione a mosaico di alleli posti sul cromosoma X è la pezzatura, detta a squama di tartaruga, del manto di alcune gatte non riscontrabile nei gatti (maschi). La femmina è eterozigote: un cromosoma X

porta il colore rosso-marrone (recessivo) e l'altro il nero (dominante). Si ha la comparsa di macchie rosso-marroni e nere sul manto bianco, colore dato da un altro gene. Le macchie rosse compaiono quando è inattivato il cromosoma X che porta il gene che conferisce il colore nero. Se non ci fosse l'inattivazione del cromosoma X, le macchie sarebbero solo nere. Se ambedue i cromosomi X fossero inattivi le macchie sarebbero solo rosse.
L'individuo chimera risulta dall'aggregazione di due zigoti o di giovani embrioni (geneticamente diversi) che portano alla formazione di un unico individuo che pertanto risulta formato da due tipi di cellule geneticamente diverse. E' il meccanismo inverso a quello che da un unico zigote porta alla formazione di due gemelli. La chimera si può formare anche per colonizzazione di un embrione da parte di poche cellule provenienti dal suo gemello geneticamente non identico. I centri trasfusionali occasionalmente individuano chimere tra i donatori di sangue.

10. *Impronta dei genitori o impronta genomica.* L'impronta dei genitori (parental imprinting or genomic imprinting) è responsabile del diverso grado di espressione dell'allele di uno stesso gene in relazione all'origine paterna o materna del cromosoma dove è localizzato il gene. Pertanto le cellule dell'individuo esprimeranno o solo l'allele materno o solo quello paterno.
L'impronta dei genitori interessa solo alcuni geni, si stima che nell'uomo siano 100-200 geni riuniti in 15-30 gruppi (cluster). Essa è stabile per una sola generazione perché è cancellata (prima/durante la gametogenesi) ed impressa nuovamente durante la gametogenesi. Si assume che l'impronta dei genitori possa essere impressa anche subito dopo la fecondazione, prima della fusione dei pronuclei. Così in un individuo, per un certo gruppo di geni, non sono (o sono meno) espressi gli alleli di origine paterna mentre sono espressi gli omologhi di origine materna. Per altri geni può avvenire il contrario (sono espressi quelli paterni e non quelli materni). Da ciò la diversità fenotipica di un individuo in relazione all'origine (paterna o materna) del cromosoma su cui sono localizzati i geni che subiscono l'impronta. L'individuo (figlio) porterà queste differenze, ma durante la sua gametogenesi, in relazione al suo sesso (es. maschio), darà a tutti i suoi cromosomi (sia quelli avuti dal padre che dalla madre) la tipica impronta dei genitori del suo sesso (maschile). Poiché l'impronta dei genitori non è trasmessa alla seconda generazione (nipoti) si ammette che essa sia esclusivamente di natura epigenetica, cioè che non venga alterata in nessuna regione la sequenza nucleotidica del DNA cromosomico. Uno dei meccanismi (forse l'unico) tramite il quale si realizza l'impronta dei genitori è la metilazione del DNA. La metilazione delle citidine è associata ad una ridotta attività di trascrizione. Tuttavia la metilazione può provocare effetti positivi sull'espressione di un dato gene, quando viene metilato il gene che codifica una proteina inibitoria del primo gene. La metilazione ha le tre proprietà teoriche richieste all'impronta dei genitori: 1) presente sui geni di un gamete, deve rimanere stabile dopo la fecondazione e durante la mitosi di tutte le cellule dell'embrione e successivamente (durante la

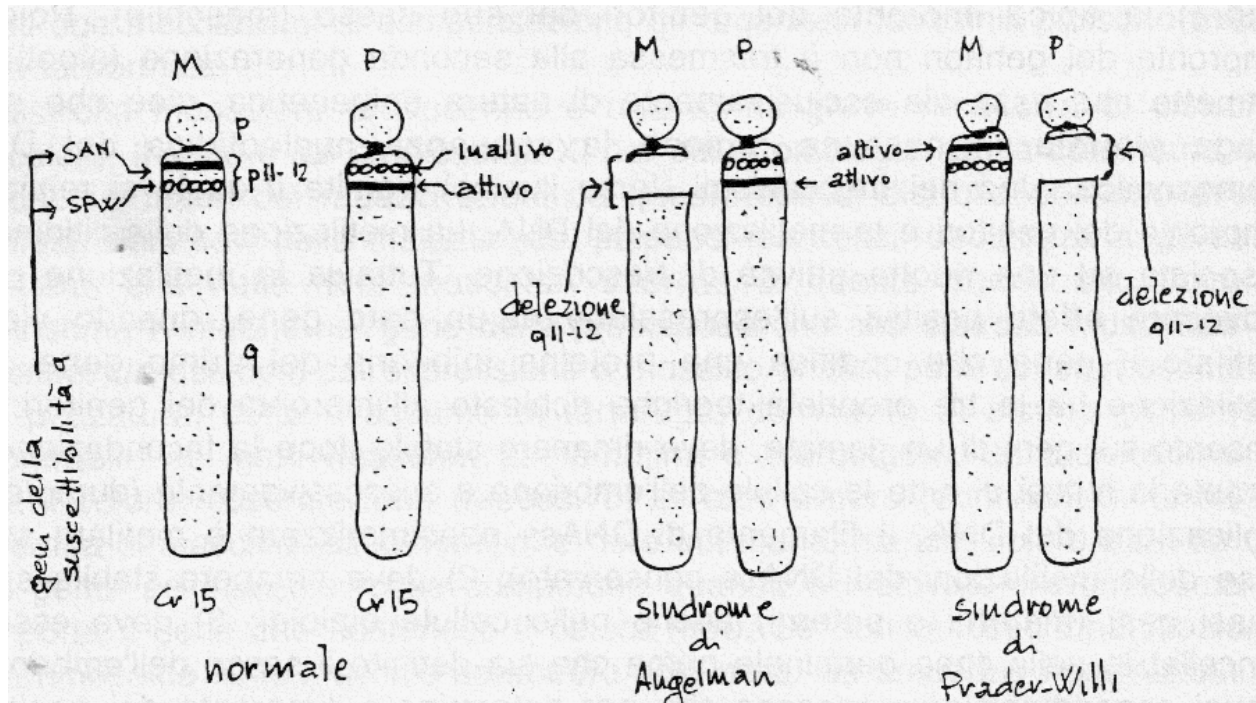



Figura D-4.
Impronta dei genitori, sindrome di Angelman (SAN) e sindrome di Prader-Willi (SPW).
I geni responsabili delle due sindromi sono localizzati molto vicini sul braccio lungo (q) del cromosoma 15 (Cr15q11-12). L'allele attivo della suscettibilità a SAN o a SPW è indicato con ---- e gli alleli inattivi a causa dell'impronta dei genitori (SAN o SPW) sono indicati con ∞∞. In figura le posizioni relative dei due geni, SAN più vicino e SPW più distante dal centromero, sono arbitrarie. Si assume che negli individui normali:

1. Il cromosoma 15 materno (M) abbia il gene della suscettibilità alla SAN espresso e normale (----). Se il gene è espresso la patologia non si manifesta. Mentre il gene della suscettibilità al SPW è normale ma inattivo (∞∞) in conseguenza dell'impronta dei genitori materna.
2. Il cromosoma 15 paterno (P) abbia una condizione opposta, il gene della suscettibilità alla SAN è normale ma è reso inattivo (∞∞) dall'impronta dei genitori, mentre quello della suscettibilità alla SPW è attivo (----)
3. Negli individui normali (Chr 15 ambedue integri) le inattivazioni delle due impronte genetiche non causano patologie perché è sufficiente l'attività di un solo allele della suscettibilità a SAN (materno) e quello della suscettibilità a SPW (paterno) per non avere la manifestazione della patologia.

Con la delezione nella zona q11-12 del Chr15 si ha la perdita di ambedue i geni, tuttavia la manifestazione patologica è diversa in relazione a quale dei due cromosomi (materno o paterno) ha subito la delezione. Quando la delezione è sul cromosoma Chr15 materno, si manifesta la SAN, perché la delezione ha rimosso l'unico allele attivo del gene SAN, essendo il gene SAN paterno integro ma inattivato dall'impronta dei genitori paterna. Quando la delezione è sul cromosoma Chr15 paterno si ha la situazione opposta e si manifesta la SPW. Invocando la presenza di un impronta dei genitori diversa nei due geni sui cromosomi 15 di origine paterna e materna si può spiegare l'insorgere di due patologie diverse in conseguenza della delezione della zona q11-12. Le due patologie dipendono da delezioni della zona q11-12 nel 100% dei casi.

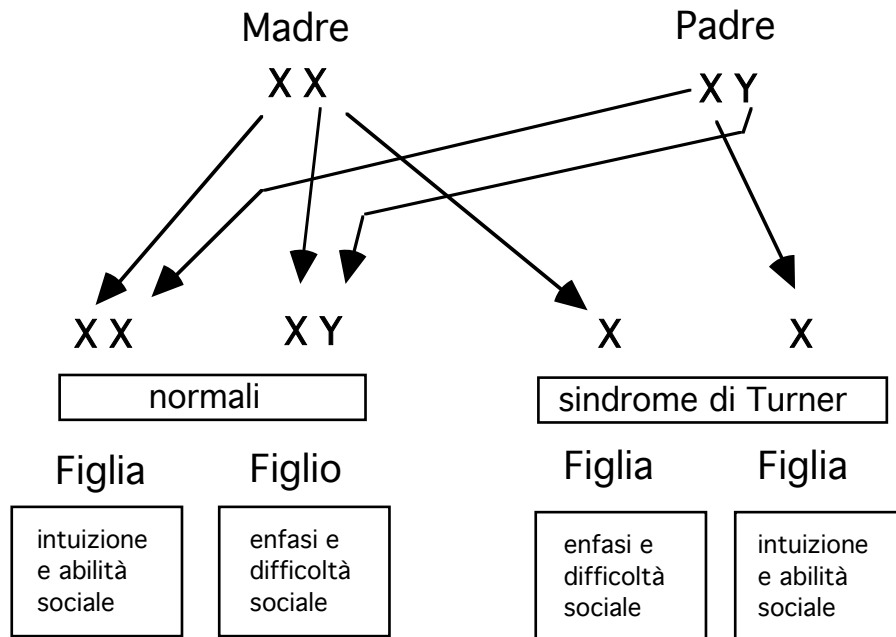



Figura D-5. Skuse ed altri (Nature 387, 705-708, 1997) hanno proposto che l'impronta del padre sul cromosoma X sia responsabile della maggiore intuizione e sensibilità sociale delle femmine, mentre l'impronta della madre sia responsabile della maggiore tendenza all'enfasi ed alle difficoltà sociali (alterazione della vita familiare, insensibilità allo stato d'animo delle persone) dell'uomo. Analizzando ragazze con un solo cromosoma X (sindrome di Turner) si è osservato che avevano intelligenza normale ma quelle che avevano ricevuto il cromosoma X materno (secondo alcuni parametri che includono quelli sopra indicati) avevano minore abilità sociale (come i maschi normali) mentre quelle che avevano ricevuto il cromosoma X paterno avevano più intuizione ed abilità sociale. Ciò sembra concordare con l'osservazione che i maschi piuttosto che le femmine sviluppano patologie come l'autismo che includono difficoltà ad interagire con gli altri. Ovviamente esistono fenocopie, cioè maschi con sensibilità sociale e femmine con difficoltà sociali a causa delle condizioni ambientali. (Gli autori in una intervista hanno spiegato scherzando che la selezione naturale avrebbe favorito i maschi con scarsa sensibilità perché dovevano uscire dalla tana per uccidere animali o simili, mentre la sensibilità sociale delle donne era utile per aspettare pazientemente e poi attrarre/calmare gli uomini se ritornavano vivi e un po' nervosi).

---

replicazione del DNA, il filamento di DNAss neosintetizzato è metilato sulla base delle metilazioni del DNAss conservato); 2) deve rimanere stabile sugli stessi geni (materni o paterni) anche nelle cellule diploidi; 3) deve essere cancellabile nella linea germinale prima che sia definito il sesso dell'embrione. Alcuni esperimenti suggeriscono che per determinare l'impronta dei genitori, oltre alla metilazione siano necessarie delle proteine che interagiscono con i gruppi metili. Per alcuni geni l'impronta dei genitori può essere persa fisiologicamente nell'adulto, probabilmente perché ha svolto la sua funzione durante l'embriogenesi e/o la formazione degli organi e nell'adulto non è più

necessario il suo controllo su quei geni. Mentre per altri geni, l'impronta dei genitori ha un ruolo anche nell'adulto, e la sua perdita causa tumori o altre patologie.
Mediante microchirurgia è possibile trasferire dalle uova fecondate i pronuclei (uno contiene il corredo cromosomico aploide della madre e l'altro quello del padre). Dopo centinaia di esperimenti di trasferimento di pronuclei di topo (Mus musculus) si è osservato che gli embrioni che avevano ricevuto sia il corredo materno che quello paterno si sviluppavano normalmente. Mentre gli embrioni che avevano ricevuto un doppio corredo aploide, solo materno (ginecogenetici) o solo paterno (androgenetici) non sopravvivevano. L'esperimento suggerisce che per un normale sviluppo embrionale non è sufficiente avere un corredo diploide ma occorre che un corredo provenga da un maschio e l'altro dalla femmina. Si suppone che durante la gametogenesi maschile e femminile i rispettivi corredi cromosomici subiscano l'impronta dei genitori e che l'impronta svolga un ruolo nello sviluppo embrionale e fetale. Casi simili sono stati osservati nell'uomo. Per degenerazione del pronucleo materno si ha la formazione di un trofoblasto, con corredo cromosomico normale esclusivamente maschile, incapace a formare un embrione. Tuttavia, questa è una prova negativa e non può escludere che in rari casi si possano avere individui ginecogenetici o androgenetici.
Nell'uomo sono stati individuati 16 geni (5 con impronta materna ed 11 paterna) riuniti in due raggruppamenti (cluster). Essi includono la regione dei geni che predispongono alle sindromi di Angelman e di Prader-Willi (figura D-4), il gene soppressore (antioncogene) del tumore di Wilms, il gene dell'insulina, quello del fattore di crescita insulina-simile, quello di un fattore di trascrizione e di altri geni.
La delezione della parte prossimale del braccio lungo (q) del cromosoma 15 mostra la presenza dell'impronta dei genitori nell'uomo. Se la delezione è sul cromosoma materno si ha (quasi invariabilmente) la sindrome di Angelman (SAN), mentre se è su quello paterno si ha (quasi invariabilmente) la sindrome di Prader-Willi (SPW) (figura D-4). Quindi si hanno fenotipi patologici diversi in relazione alla provenienza (paterna o materna) del cromosoma, sul quale sono localizzati i geni mutati responsabili delle due patologie.
La malattia di Huntington è una malattia ereditaria autosomica dominante (locus 4p) che provoca gravi disturbi neurologici. Gli individui che ricevono il gene mutato dal padre mostrano i sintomi durante l'adolescenza, mentre in quelli che lo ricevono dalla madre i sintomi appaiono molto dopo, durante l'età adulta (30-50 anni). Questa differenza di penetranza della malattia è imputata alla impronta dei genitori.
L'impronta dei genitori è stata osservata anche in altre patologie umane: atassia spinocerebellare, distrofia miotonica e tumore di Wilms.
Si riteneva che l'impronta dei genitori interessasse solo gli autosomi, invece alcune recenti osservazioni suggeriscono che essa sia presente anche sul cromosoma-X ed influenzi il comportamento maschile e femminile (figura D-5).

Il termine impronta dei genitori o genomica non deve essere confuso con i termini: impronta genetica o impronta del DNA (traduzione di DNA fingerprint = impronta digitale del DNA) che sono utilizzati per indicare i dati genetici sufficienti ad individuare un individuo mediante l'analisi di una serie di sequenze ripetute del DNA (capitolo 3).

11. *Cromosomi mitocondriali.*
I mitocondri hanno un proprio genoma costituito da 5-10 copie di un cromosoma di DNAds circolare privo di istoni (migliaia di cromosomi/cellula). Il DNAds mitocondriale ha 16.569b e codifica 37 geni per rRNA, tRNA e 13 subunità di enzimi respiratori. Il DNA mitocondriale è trasmesso esclusivamente per via materna ed è una componente molecolare dell'eredità esclusivamente materna.
Esistono patologie causate da mutazioni del DNA mitocondriale come un tipo di cecità: la neuropatia ottica di Leber che è caratterizzata da un improvvisa irreversibile perdita della vista nei giovani adulti. Di questa malattia sono responsabili indipendentemente due geni diversi, mutati in maniera puntiforme. I loro prodotti sono coinvolti nella fosforilazione ossidativa. Tutti i tessuti sono interessati, incluse le gonadi, per cui la madre trasmette la malattia alla prole (maschi e femmine). Esistono anche miopatie causate da delezioni o da duplicazioni del DNA mitocondriale che possono essere localizzate solo in alcuni tessuti. E' importante stabilire se le ovaie siano risparmiate per poter escludere la possibilità di trasmissione.

Alcune considerazioni sugli aspetti molecolari della dominanza e della recessività dei caratteri genetici

La dominanza è caratterizzata da una prevalenza dell'attività del prodotto genico (carattere dominante) su quella del prodotto genico dell'allele recessivo. La prevalenza dell'attività biologica del prodotto genico di un allele rispetto all'altro può essere determinata mediante i seguenti meccanismi:
a) il gene esprime monomeri ed il prodotto genico di un allele è biologicamente più attivo (dominate) dell'altro (recessivo). Se il prodotto genico è una proteina enzimatica la maggiore attività può risultare da una maggiore affinità verso il substrato e/o maggiore attività di catalisi e/o da una maggiore concentrazione allo stato stazionario di proteina nella cellula. L'azione dell'enzima dominante prevale su quella dell'enzima recessivo perché utilizza più substrato e forma più prodotto dell'enzima recessivo.
b) se il gene esprime un protomero, cioè una proteina subunità di un oligomero, in condizione di eterozigosi si formeranno oligomeri misti. Valgono ancora le considerazioni fatte in a). Infatti le subunità cataliticamente più attive prevarranno su quelle meno attive anche se appartenenti alla stessa proteina oligomerica. Se un allele produce una proteina in maggiore quantità ci sarà la

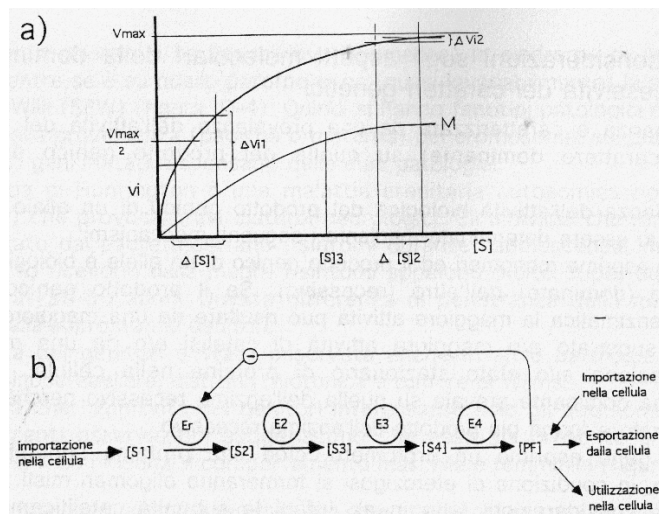



Figura D-6.Regolazione da substrato e da effettore di enzimi di una via metabolica.
a) curva di saturazione da substrato di un enzima regolato solo da substrato. Sulle ascisse la velocità iniziale (*vi*) sulle ordinate la [S]. La *vi* con valori vicini a 1/2Vmax è regolata efficientemente da piccole variazioni di [S]. Se *vi* ha valori inferiori ad 1/2Vmax, la regolazione da [S] incrementa la sua efficienza (maggiore variazione della *vi* per gli stessi valori di Δ[S]), tuttavia la *vi* è minore rispetto a 1/2Vmax; mentre quando *vi* ha valori vicini a Vmax è scarsamente regolabile per le stesse variazioni di [S]. Quindi con *vi*=1/2Vmax si ha la maggiore velocità di catalisi associata alla più efficiente regolazione per variazioni di [S].
b) schema della regolazione a inibizione retrogada della via metabolica lineare. Er, enzima regolato dal prodotto finale (PF) che è effettore negativo; E_2, E_3 ed E_4 enzimi regolati solo da substrato. Er ha una attività totale inferiore a quella degli enzimi non regolati e ciò in genere dipende da una sua minore concentrazione cellulare. Risulta così che l'attività catalitica massima di Er è la velocità massima del flusso metabolico, cioè di sintesi di PF, (l'attività di Er è detta: rate limiting step = passo che limita la velocità). Inoltre Er ha un turnover più rapido di quello degli altri enzimi in quanto la sua concentrazione può essere aggiustata rapidamente al fine di poter variare opportunamente il flusso della via metabolica. La reazione catalizzata da Er è continuamente mantenuta fuori dall'equilibrio e tende a raggiungerlo nella direzione del prodotto, mentre le altre reazioni hanno reagenti e prodotti in concentrazioni uguali a quelle dell'equilibrio termodinamico. Anche quando Er è inibito al massimo rimane un minimo ma continuo flusso metabolico, per cui le reazioni non regolate sono sempre in equilibrio allo stato stazionario (non termodinamico perché le reazioni procedono con trasferimento di materia). Queste caratteristiche della via metabolica permettono ad Er di regolarla efficientemente. La maggiore attività catalitica degli enzimi non regolati rispetto a quella di Er, permette ad Er di aggiustare rapidamente la velocità del flusso metabolico in accordo con la sua attività catalitica. Ad esempio, quando viene ridotta l'inibizione di Er per riduzione della concentrazione di PF, la maggiore produzione di S_2, mediante le reazioni intermedie, viene rapidamente convertita in PF con pochissima perdita di energia perché le reazioni sono all'equilibrio allo stato stazionario. Il meccanismo funziona rapidamente anche in senso contrario. Se l'attività catalitica degli enzimi E_2, E_3 ed E_4 non fosse potenzialmente superiore a quella di Er, gli incrementi di attività di Er sarebbero limitati dall'attività di questi enzimi. All'interno di questo limite massimo, gli enzimi regolati sono anche responsabili dell'entità del flusso metabolico in relazione alla concentrazione allo stato stazionario di PF che a sua volta dipende dall'importazione, dall'esportazione e dall'utilizzazione di PF nella cellula. Per questo gli enzimi regolati sono detti anche "pace maker" (fattori del ritmo) della via metabolica. Il normale flusso metabolico e la sua corretta regolazione dipendono dal continuo apporto nella cellula del primo substrato (S_1). Se questo è insufficiente, la via metabolica e la sua regolazione sono inefficienti.
Assumiamo che l'enzima E_3 sia mutato in condizioni di eterozigosi ed M sia la sua curva di saturazione da substrato. A causa della carenza di attività di E_3, la concentrazione allo stato stazionario di S_3 (substrato di E_3) aumenta di circa tre volte causando un incremento della *vi* di E_3 che compensa in parte l'attività persa in conseguenza della mutazione. Tuttavia la mutazione causa una perdita nella potenzialità della capacità di sintesi di PF, perché l'enzima E_3 è già in condizioni di Vmax e la sua *vi* non può essere ulteriormente incrementata in caso di richiesta di una maggiore sintesi di PF. La mutazione di E_3 sarebbe dominante se la recuperata attività di E_3 fosse insufficiente per le normali richieste di PF o se S3 in alte concentrazioni fosse tossico per le cellule dell'organismo.

prevalenza numerica di questa subunità nella totalità degli oligomeri. Questa dominanza può portare la concentrazione allo stato stazionario dei prodotti di reazione a livelli più alti e ciò può influenzare attraverso meccanismi di regolazione allosterica altri enzimi/proteine ed in questo modo portare alla manifestazione delle caratteristiche fenotipiche della dominanza.
Se negli oligomeri misti, un protomero dominante negativo prodotto da un allele inibisce l'attività del protomero prodotto dall'altro allele si ha la prevalenza del primo sul secondo prodotto genico.
Ci si può domandare come un enzima di una via metabolica che può esistere espresso dal relativo gene in tre condizioni, omozigosi dominante, eterozigosi, ed omozigosi recessiva, possa svolgere egualmente bene la sua attività biologica quando nelle tre condizioni genetiche la sua attività catalitica abbia valori molto diversi. Assumiamo che l'allele recessivo codifichi un enzima che catalizza la stessa reazione, ma abbia minore affinità per il substrato e/o minore attività catalitica. Se in condizioni di eterozigosi l'attività della proteina enzimatica è sufficiente ad operare la catalisi per le normali esigenze della cellula, in condizioni di omozigosi dominante, la maggiore attività dovrebbe essere eccessiva e creare problemi di accumulo di metaboliti non necessari alla cellula con spreco di energia e di composti. Mentre in condizioni di omozigosi recessiva si dovrebbe avere una condizione di insufficiente attività.
Una risposta si ottiene considerando i meccanismi che regolano la via metabolica. In una via metabolica gli enzimi sono saturati al 50% dai rispettivi substrati e funzionano a circa la metà della loro velocità massima (Vmax).
Questa condizione permette la più alta velocità associata alla migliore (più pronta) regolazione per relativamente piccole variazioni positive e negative della concentrazione del substrato. Ciò è facilmente verificabile osservando la curva iperbolica data dalla velocità iniziale di catalisi in dipendenza dalla saturazione dell'enzima da parte del substrato (figura D-6a). Se gli enzimi per le normali esigenze metaboliche basali della cellula dovessero funzionare alla velocità massima (enzimi saturati dal substrato), per la presenza del plateau, non potrebbero incrementare la loro attività quando la cellula, richiedendo una maggiore attività della via metabolica, incrementasse l'apporto del primo substrato. E per la stessa ragione, neppure ridurre la velocità di reazione mediante le riduzioni di concentrazione del substrato con le quali la cellula opera la regolazione.
In una via metabolica, quando aumenta l'apporto del primo substrato, automaticamente aumenta l'attività catalitica del primo enzima, di conseguenza aumenta anche la concentrazione del suo prodotto di reazione che è substrato del secondo enzima della via metabolica. Il secondo enzima viene attivato dall'azione del primo, perché esposto ad una maggiore concentrazione di substrato. Nello stesso modo il secondo enzima attiva il terzo e così avanti fino al prodotto finale che è richiesto per la fisiologia cellulare del momento. Lo stesso meccanismo procede anche in senso inverso, quando diminuisce la concentrazione del primo substrato, diminuisce anche l'attività del primo enzima e quindi di tutti gli altri enzimi della stessa via

metabolica. Questo tipo di regolazione degli enzimi e quindi delle vie metaboliche è detta <u>regolazione da substrato</u>.
Oscillazioni delle concentrazioni degli intermedi delle vie metaboliche avvengono continuamente nella cellula in risposta alle variazione dell'apporto dei primi metaboliti da parte del sangue in conseguenza delle variazioni dell'alimentazione, stato di digiuno, stato ormonale, esercizio fisico o dell'utilizzazione dei metaboliti prodotti finali da parte delle cellule dei vari organi.
L'espressione dei geni che codificano gli enzimi di una via metabolica appare regolata in modo che le concentrazioni cellulari di tutti gli enzimi siano tali da garantire sia le alte che le basse attività metaboliche per la normale fisiologia cellulare. Si assume che queste concentrazioni siano vicine a quelle della semisaturazione da parte dei rispettivi substrati e diano valori di velocità iniziale intorno alla metà della Vmax.
La regolazione da substrato delle vie metaboliche è a sua volta controllata da un meccanismo di regolazione detto a <u>inibizione retrograda</u> (feedback negativo) che ha la capacità di opporsi entro certi limiti alle variazioni di apporto del primo substrato (figura D-6b). Con questo meccanismo il prodotto finale (PF) della via metabolica è effettore negativo di un enzima della stessa via. L'enzima, detto regolato, è il primo della via o comunque è in una posizione da poterla regolare efficientemente. Risulta così che la concentrazione allo stato stazionario (CSS) del prodotto finale (quello utilizzabile solo dalla cellula che lo sintetizza (es. ATP) o dalle cellule di più tessuti di uno stesso organismo, es. glucosio, acidi grassi) si autoregola, controllando l'attività della via metabolica. Quando la cellula utilizza una maggiore quantità di prodotto finale la CSS di questo si abbassa, l'inibizione sul primo enzima diminuisce e ciò provoca un incremento della produzione del suo prodotto di reazione che aumenta in CSS. Questo aumento di CSS, per regolazione da substrato, attiva il secondo enzima e così via fino al prodotto finale. Il meccanismo procede efficientemente anche in senso inverso quando la cellula riducendo l'utilizzazione del prodotto finale, ne fa aumentare la CSS. L'aumento della CSS del PF inibisce l'enzima regolato causando la riduzione della sintesi del prodotto finale stesso fino a quando la sua CSS è tale da accordare la velocità di sintesi del prodotto finale con quella della sua utilizzazione.
Il meccanismo a inibizione retrogada opera efficientemente anche quando l'utilizzazione del prodotto finale rimane costante mentre si hanno variazioni (incrementi o decrementi) di concentrazione del primo substrato. Oppure quando il prodotto finale, aumentando in concentrazione perché importato dal sangue nella cellula, automaticamente limita la sua produzione endogena (es. il colesterolo della dieta inibisce la sua sintesi endogena).

La regolazione da substrato ha l'importante funzione di regolare l'attività di una via metabolica (sintesi del prodotto finale) in relazione diretta alla disponibilità del primo substrato, coordinando tra loro le varie reazioni enzimatiche. La regolazione a inibizione

retrogada ha la funzione di rendere la regolazione da substrato dipendente dall'apporto e dall'utilizzazione del prodotto finale in relazione alle attività fisiologiche della cellula e/o dell'organismo. Ambedue i meccanismi di regolazione possono funzionare se nella cellula c'è una continua disponibilità (mai annullata completamente) del primo substrato.

La regolazione da substrato e quella a inibizione retrogada, governando la via metabolica in relazione alle necessità fisiologiche del prodotto finale, permettono di compensare le possibili differenze di attività specifica e/o di concentrazione degli enzimi che possono risultare dalla normale variabilità genetica degli alleli legata alla dominanza, alla recessività ed al polimorfismo.
La regolazione da substrato causa anche un altro tipo di compensazione che agisce sulla stabilità degli enzimi. La maggiore saturazione dell'enzima, causata dalla regolazione da substrato, risulta oltre che in una maggiore attività anche in una maggiore stabilità dell'enzima. L'enzima complessato con il substrato è più stabile perché ha meno tendenza a denaturarsi ed a subire l'attacco delle proteasi cellulari. L'incremento della stabilità causa un incremento della concentrazione dell'enzima anche se la sua sintesi rimane costante. Ciò risulta in un ulteriore incremento dell'attività di quell'enzima nella cellula che contribuisce a compensare la minore attività catalitica degli enzimi cataliticamente meno attivi e/o meno concentrati (un enzima più saturo di substrato è più attivo e più stabile). Il meccanismo funziona anche in senso opposto, quando la concentrazione del primo substrato decresce, aumenta il numero degli enzimi liberi dal substrato, i quali sono più facilmente attaccati dai sistemi proteolitici. Di conseguenza la concentrazione dell'enzima decresce.

Ipotesi sui meccanismi molecolari di compensazione delle attività enzimatiche che per cause genetiche risultano carenti o eccessive per le normali funzioni metaboliche.

Se un enzima, ad esempio E_3, della via metabolica in figura D-6b per cause di una mutazione risulta avere meno attività totale per cellula dell'enzima (E_2) ad esso precedente nella via metabolica, si ha l'incremento della CSS di S_3 (prodotto di E_2 e substrato di E_3) che provocherà l'incremento dell'attività catalitica di E_3. La CSS di S_3 continuerà a salire fino a quando nella cellula l'attività catalitica totale di E_3 non risulterà uguale a quella di E_2.
La regolazione da substrato compensa la minore concentrazione o attività di E_3 saturando di substrato un maggior numero di molecole di E_3. Egualmente, se per cause genetiche nella cellula l'attività catalitica E_3 fosse maggiore di quella di E_2, la concentrazione di S_3 si abbasserebbe e di conseguenza anche l'attività catalitica di E_3, mentre la reazione E_2, essendo all'equilibrio allo stato stazionario, verrebbe incrementata per sottrazione di prodotto.

In alcuni casi la compensazione operata dalla regolazione da substrato può essere insufficiente ed allora interviene la regolazione da effettore con il meccanismo a inibizione retrogada.
Per spiegare il meccanismo facciamo una ipotesi:
- il primo enzima (Er) della via metabolica è regolato dal prodotto finale (PF).
-E_3 ha nella cellula una attività catalitica totale inferiore a quella di E_2 perché cataliticamente meno attivo e/o meno concentrato.
-la CSS di S_1 non può essere incrementata perché non può essere incrementato il flusso di entrata di S_1 nella cellula .
In queste condizioni, inizialmente la CSS di PF si abbasserà provocando la riduzione dell'inibizione di Er, il quale essendo un enzima regolato da effettore, incrementerà la propria attività catalitica anche se la concentrazione di S_1 rimane invariata. L'attivazione di Er incrementerà la CSS di S_2 che per regolazione da substrato farà aumentare l'attività di E_2 che con lo stesso meccanismo farà aumentare l'attività di E_3 e cosi avanti fino a quando PF sarà sintetizzato nella quantità necessaria alla cellula.
Le attività catalitiche degli enzimi non regolati da effettore che sono più alte di quanto occorra al normale metabolismo e che porterebbero a sintetizzare quantità eccessive di PF non possono essere sempre compensate con la sola regolazione da substrato. Ciò è possibile con la regolazione da effettore perché l'eccesso di sintesi fa incrementare la CSS di PF che inibisce Er e quindi rallenta tutto il flusso metabolico fino a quando la velocità sintesi di PF eguaglia quella della sua utilizzazione. L'eccesso di sintesi/concentrazione di un enzima denuncia una incongruenza, cioè che la regolazione dell'espressione dei relativi geni non sia perfettamente calibrata con la funzione che gli enzimi devono svolgere nelle cellule, fino all'estremo di esprimere proteine che non hanno alcuna funzione fisiologica. Dati ottenuti con la distruzione di singoli geni (gene knockout) indicano che alcuni geni sono espressi in modo costitutivo nelle cellule di più organi, ma in alcuni organi le proteine espresse non svolgono alcuna funzione (l'organo appare morfologicamente e metabolicamente normale anche in assenza della proteina). Il dato è interpretato assumendo che la regolazione genica sia energeticamente molto costosa perché in genere richiede la sintesi di più proteine per regolarne una, pertanto se la proteina codificata dal gene non è tossica per la cellula, essa può essere sintetizzata anche in eccesso o anche in cellule che non necessitano dell'attività della proteina. Se per esempio consideriamo un enzima che è espresso in più organi aventi livelli diversi di attività metaboliche, si tende ad assumere che la concentrazione dell'enzima nelle cellule dei vari organi debba essere diversa e di conseguenza anche il grado di espressione (trascrizione e traduzione) del relativo gene dovrebbe essere diversa. Tuttavia ciò non sembra verificarsi per tutti i geni. Si ipotizza che modulare l'espressione di uno stesso gene in modo diverso in tutti i tessuti richiederebbe molta energia (sintesi di specifici fattori di trascrizione) e risulterebbe più conveniente esprimere il gene in modo costitutivo e poi regolare più finemente a livello del fenotipo, cioè regolare

l'attività molecolare della proteina in modo diverso nelle cellule dei vari tessuti. Come discusso prima per gli enzimi, ciò può avvenire mediante la regolazione da substrato e da effettore.
Se il flusso nelle cellule del 1° substrato (S_1) ha piccole oscillazioni, esse possono essere compensate con variazioni dell'attività dell'enzima regolato (Er) mediante la sua regolazione da substrato come visto per gli enzimi non regolati. Quando nella cellula la CSS di S_1 è costantemente più alta (ad es. per incremento della [S_1] ematica) o più bassa (es. S_1 è utilizzato in altre vie metaboliche) di quella occorrente per sintetizzare PF nelle quantità necessarie alla cellula, la compensazione è operata con la regolazione a inibizione retrogada da effettore (PF). La compensazione da effettore appare necessaria anche quando alcuni metaboliti o cofattori della via metabolica (es. ATP/ADP/AMP, NAD/NADH, NADP/NADPH, ecc.) partecipano ad altre vie come substrati e/o effettori di regolazione ed hanno la funzione di coordinare più vie metaboliche. Un'alta CSS di ATP indica uno stato energetico ottimale per la cellula (sono attivate le vie di sintesi), bassa CSS dell'ATP ed alta CSS AMP indicano uno stato di carenza energetica, le vie di sintesi sono inibite ed attivate quelle del metabolismo energetico per la produzione di ATP. Le CSS di ATP ed AMP possono variare entro limiti molto precisi, altrimenti gran parte del metabolismo cellulare sarebbe sconvolto. A sostegno di questa ipotesi sta l'osservazione che le mutazioni dei geni che codificano enzimi regolati da effettore a inibizione retrogada di vie metaboliche sintetiche sono dominanti e quelle dei geni che codificano enzimi regolati da substrato che fanno parte di vie cataboliche sono recessive. Per gli enzimi regolati da effettore, ma non per quelli non regolati (regolati da substrato), la perdita di metà della concentrazione è sufficiente a rendere inefficiente la regolazione della via metabolica.
La compensazione di attività enzimatiche carenti ha un costo consistente nella perdita di potenzialità di incremento del flusso metabolico.
Assumiamo che il patrimonio genetico di un individuo codifichi un enzima (enzima-A) regolato solo da substrato e con attività catalitica totale per cellula minore degli altri enzimi appartenenti alla stessa via metabolica. La regolazione da substrato compenserà la minore attività catalitica saturando maggiormente l'enzima-A con un incremento della concentrazione del relativo substrato.
Tuttavia ciò limita la potenzialità dell'incremento di attività dell'enzima-A, perché in condizioni metaboliche di base, la frazione di molecole di enzima-A impegnate nella catalisi risulta maggiore di quella degli altri enzimi della via metabolica. Infatti, quando per le normali esigenze fisiologiche il flusso metabolico deve essere incrementato al massimo, l'enzima-A potrà incrementare la propria attività catalitica di una quantità inferiore rispetto a quella degli altri enzimi della stessa via metabolica, limitando così il flusso metabolico massimo al massimo valore della sua attività catalitica (geneticamente più bassa).
Un analogo ragionamento può essere fatto quando è utilizzata la regolazione a inibizione retrograda per compensare l'attività catalitica di uno o più enzimi

aventi carente attività catalitica. L'enzima regolato per compensare le carenze catalitiche della sua stessa molecola o di altri enzimi della via metabolica, invece di essere inibito al 50%, risulta inibito solo al 30% (cioè attivo al 70%). Il 20% in più di attività dell'enzima serve a compensare una minore attività sua e/o di altri enzimi al fine di sintetizzare il PF (es. ATP) per le normali esigenze di base della cellula (es. fibra muscolare in riposo). Quando la cellula richiede una maggiore sintesi di prodotto finale (es. sforzo muscolare), l'enzima regolato ha a disposizione solo il 30% della sua attività per rispondere alla maggiore richiesta metabolica di PF. Mentre quando lo stesso enzima regolato è naturalmente più attivo (perché espresso da un altro allele dello stesso gene) per alimentare lo stesso flusso metabolico è inibito solo del 50%. Pertanto l'incremento di attività che può essere richiesto è pari al 50% dell'attività catalitica totale.
Individui normali, perché geneticamente eterogenei, hanno necessità diverse di compensare le possibili differenze di attività delle proprie proteine coinvolte nelle vie metaboliche ed in tutte le altre funzioni cellulari ed extracellulari. Da ciò scaturisce una diversa potenzialità fisiologica determinata dalla costituzione genetica di ogni individuo.
Queste diverse potenzialità sono uno dei fattori responsabili delle diverse prestazioni di individui normali geneticamente diversi che si cimentano nella stessa attività (sportiva, intellettuale, ecc.). Gli altri fattori sono l'ambiente e l'alimentazione.
I geni, controllando la quantità e qualità delle macromolecole, controllano anche la quantità e qualità (differenziamento) delle cellule di un dato tessuto o organo. Ad esempio i muscoli che svolgono contrazioni lente e durature (postura, corsa di fondo) hanno prevalentemente fibre di tipo I che sono ricche di mitocondri e di mioglobina perché producono ATP con il metabolismo aerobico (glicolisi e ciclo di Krebs). I muscoli che svolgono contrazioni veloci e brevi (movimenti degli occhi, corsa veloce) hanno prevalentemente fibre di tipo IIA (simili alle fibre I ma più rapide nelle contrazioni) e IIB. Le fibre IIB utilizzano esclusivamente il metabolismo anaerobico (glicolisi), sono povere di mitocondri e di mioglobina, ma hanno una maggiore concentrazione di enzimi glicolitici, sono poste nella parte esterna del muscolo meno vascolarizzata e sono responsabili delle contrazioni rapide anaerobiche.
Il fenotipo della normale costituzione genetica di un muscolo può essere modificato con l'allenamento. Con l'allenamento alla corsa di fondo (10.000 metri, ed oltre) nei muscoli delle gambe (che si contraggono per quello sforzo) le fibre di tipo I aumentano in numero e dimensioni (si arricchiscono di mitocondri, mioglobina ed enzimi) mentre si riducono in numero le fibre di tipo IIB che acquistano le caratteristiche biochimiche (metabolismo aerobico) delle fibre di tipo IIA. Con l'allenamento alla corsa veloce (100-200 metri) i muscoli delle gambe incrementano il numero delle fibre di tipo IIA. Atleti geneticamente meno dotati con opportuni allenamenti (incrementando la dotazione di enzimi ed il numero di fibre) possono ottenere fenotipi muscolari superiori a quelli di individui geneticamente meglio dotati. La storia della cultura, la didattica e lo

studio indicano che anche le qualità intellettuali possono essere incrementate con opportuni e costanti allenamenti. In genere la costanza è caratteristica di successo.
L'esempio della compensazione della variabilità genetica degli enzimi di una via metabolica può essere generalizzato ad altri sistemi costituiti da più macromolecole. Le funzioni cellulari (metaboliche, di regolazione, ecc.) si realizzano attraverso una serie di passaggi costituiti da reazioni covalenti o di associazione per interazioni secondarie tra molecole. Questi sistemi a più passaggi sono dotati di meccanismi automatici di regolazione aventi (salvo poche eccezioni) una gerarchia di funzione crescente in questo ordine: da substrato, da effettore, da ormone/fattore di crescita, da sistema nervoso e da controllo genetico (appendice B).
Il doppio sistema di regolazione (da substrato e da effettore) è capace di mantenere la normale sintesi del prodotto finale di una via metabolica, se le carenze o gli eccessi di attività enzimatiche provocate da cause genetiche si mantengono entro certi limiti. Se i due sistemi di regolazione sono insufficienti a compensare l'alterazione genetica di uno o più enzimi di una via metabolica, la cellula si trova in uno stato patologico, l'allele responsabile e la proteina da esso codificata sono detti patologici e la manifestazione fenotipica alterata è il sintomo clinico (appendice E). La causa della patologia (eziologia) è la mutazione dell'allele che causa l'alterazione della attività molecolare e quindi anche della funzione cellulare della proteina codificata. Il meccanismo genetico-molecolare che porta alla patologia con la manifestazione dei sintomi è la patogenesi.
Nelle patologie poligeniche lo stato patologico può instaurarsi in conseguenza di una combinazione di alleli normali di geni diversi. Ad esempio: la presenza in una stessa via metabolica di più enzimi (codificati da alleli normali) aventi carente attività catalitica non compensata dalla regolazione da substrato e da effettore causa l’instaurarsi della patologia per carenza di prodotto finale. Se ad esempio nella stessa via metabolica fosse presente solo uno degli enzimi carenti in attività catalitica si attuerebbe una compensazione capace di produrre il prodotto finale sufficiente a non far instaurare la patologia. Questo sembra essere un aspetto generale degli stati patologici poligenici, cioè il superamento della soglia della capacità dei sistemi di regolazione di mantenere l'attività di una via metabolica in accordo con le necessità della cellula e dell'organismo nei vari stati metabolici (digiuno, alimentazione, riposo e sforzo muscolare, stress, ecc.). Superata la capacità di compensazione dei sistemi di regolazione si ha l'insorgere di stati patologici per carenza (od eccesso) di normali attività biologiche.
Queste carenze o eccessi possono manifestarsi sempre o anche in un solo stato metabolico. Ad esempio, alcuni stati patologici o prepatologici si possono manifestare durante lo sforzo muscolare ma non durante il riposo, altri solo durante il digiuno prolungato, altri dopo l'alimentazione. Quanto detto sopra suggerisce che il confine tra fisiologico e patologico possa essere anche molto labile e legato alla genetica dell'individuo.

La carenza o un eccesso patologico di PF possono anche risultare da un carente o eccessivo apporto del primo substrato che i sistemi di regolazione cellulare non riescono a compensare. Come si può compensare una mancanza prolungata di cibo? Ma anche l'eccesso nel tempo diviene dannoso perché il nostro organismo si è selezionato per immagazzinare al massimo gli alimenti introdotti nell'organismo. Per cui se l'alimentazione è eccessiva si finisce per immagazzinare un eccesso di grassi, che sono la forma di riserva di ogni tipo di alimento assunto in eccesso e possono essere immagazzinati in grandi quantità. Tuttavia l'eccesso di grassi porta a problemi di vario tipo (circolatori, di compressione della colonna vertebrale, ecc.).

Alcune considerazioni sul polimorfismo genetico della specie umana

Il polimorfismo genetico degli individui, ed in particolare quello SNP, data la sua alta frequenza nel genoma, è importante per la sopravvivenza di una specie perché permette a questa di rispondere ai cambiamenti rapidi dell'ambiente che selezionano positivamente alcuni individui, mentre condizionano negativamente altri che avevano vita normale nelle precedenti condizioni ambientali. Il mantenimento nella popolazione umana di alleli meno atti alla vita del momento e la possibilità di produrne altri in conseguenza della suscettibilità del genoma alle mutazioni hanno un costo che è pagato dal sacrificio degli individui portatori di malattie genetiche trasmesse o prodotte da mutazioni somatiche ed anche dagli individui sani ma meno adatti (portatori di alleli con prodotto genico meno adatto: troppo o troppo poco attivo, più o meno specifico, più o meno stabile, ecc.) a fare fronte all'ambiente (fisico, chimico, biologico e culturale) ed all'alimentazione del periodo in cui vivono. Questi aspetti della genetica indicano come gli individui di una stessa specie siano una unità molto eterogenea anche a livello molecolare, e ciò con l'apparente finalità di conservare la specie anche quando l'ambiente diventa sfavorevole per la maggior parte degli individui ed i portatori di particolari alleli non favoriti nel precedente ambiente lo divengono nel nuovo.
I cosiddetti "inadeguati" alla vita dell'inizio di questo secolo possono avere il genoma che avrebbe permesso loro di conseguire il successo in epoche passate e/o di conseguirlo in epoche future. La conoscenza del polimorfismo genetico degli individui di una stessa specie permette di capire come una stessa molecola naturale o sintetica proveniente dall'ambiente (inalata, assunta con gli alimenti o somministrata come farmaco) possa provocare influenze di tipo ed entità diverse in individui diversi. Per mostrare a livello molecolare come individui diversi rispondono in maniera diversa all'azione della stessa molecola, facciamo l'ipotesi che la molecola sia un farmaco. Alcuni individui (curabili) rispondono positivamente all'azione del farmaco, altri (non curabili con lo stesso farmaco) non subiscono alcuna influenza positiva e tra gli individui di questi due gruppi possono esserci individui che subiscono ed altri che non subiscono gli effetti negativi non voluti (effetti collaterali) del farmaco (figura D-7).

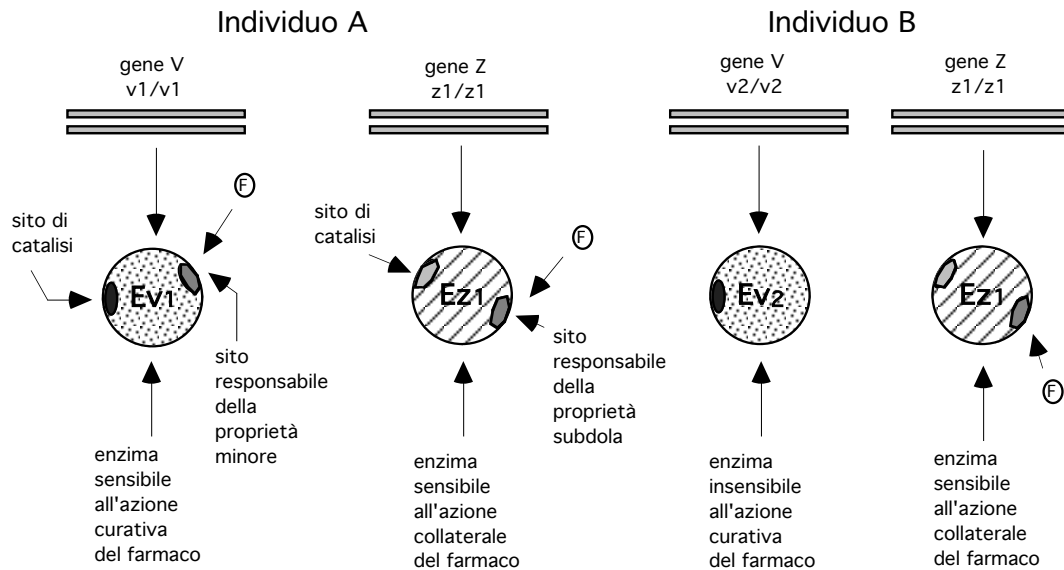



Figura D-7. Polimorfismo genetico e sensibilità ad un farmaco
Nello schema sono considerati due geni (V ed Z) di due individui (A e B) normali. Il gene V esiste in due forme alleliche v1 e v2 che codificano rispettivamente due isoforme Ev1 e Ev2 dell'enzima Ev; il gene Z esiste in una forma allelica z1 che codifica l'enzima Ez1.
Ambedue gli individui A e B sono omozigoti z1/z1, l'individuo A è omozigote v1/v1 e l'individuo B è omozigote v2/v2.
Gli alleli v1e v2 differiscono per una singola base nucleotidica e sintetizzano due isoforme dell'enzima Ev, che differiscono per un solo aminoacido posto sulla superficie dell'enzima stesso. I due enzimi hanno le stesse proprietà molecolari (attività di catalisi, regolazione e stabilità), differiscono per una proprietà minore associata all'allelle v1 ma non all'allele v2. L'enzima Ev1 (e non l'enzima Ev2) ha la capacità di associare il farmaco F che inibisce la sua attività catalitica. Ciò è importante perché il farmaco può curare una patologia causata da eccessi di attività dell'enzima Ev. Pertanto l'individuo A, ma non quello B, può essere curato con il farmaco F.
Anche l'enzima Ez1 può legare il farmaco F e da questo viene inibito causando effetti collaterali non desiderati senza portare alcun beneficio, essendo l'enzima Ez1 non coinvolto nella patologia curata dal farmaco F.

Un farmaco per poter agire deve associarsi con un certo grado di specificità ad un componente cellulare, in genere una proteina, inibendola o attivandola, per modificare la fisiologia cellulare al fine di favorire la guarigione del paziente.
Per legarsi alla proteina esso formerà dei legami deboli con una zona della superficie della proteina e cioè con dei residui aminoacidici.
Assumiamo che due pazienti possiedano alleli diversi codificanti la stessa proteina enzimatica la cui attività deve essere inibita per favorire la loro guarigione. Le due proteine hanno le stessa attività catalitica e di regolazione, tuttavia esse differiscono sulla loro superficie per uno (o pochi) residui

aminoacidici che non alterano l'attività molecolare né la funzione cellulare della proteina.
Tuttavia questa differenza fa si che una proteina possa associare il farmaco (proprietà minore/subdola) e che l'altra lo leghi con minore affinità o non lo leghi affatto. In questo secondo caso il farmaco agirà meno o non agirà del tutto (perché riuscirà a legare un numero scarso/nullo di molecole proteiche).
Nello stesso modo si può spiegare l'insorgenza di effetti collaterali di un farmaco. Assumiamo che gli stessi pazienti sopra indicati abbiano un gene che esprime una proteina, normale e non coinvolta con la patologia da curare, anche essa capace di legare il farmaco. Il farmaco inibisce (o attiva) l'azione di questa proteina, ciò causa alterazioni al normale funzionamento di uno o più organi e quindi la comparsa degli effetti collaterali. Cioè effetti nocivi e non utili alla cura della malattia.
Assumendo che i geni delle due proteine non siano associati (es. posti su cromosomi diversi), in relazione ai due tipi di proteine possedute possono esistere tre popolazioni di individui: quelli che dal farmaco ricevono solo gli effetti benefici, quelli che ricevono solo gli effetti collaterali, quelli che ricevono ambedue gli effetti o nessun effetto. Le combinazioni sono più complesse quando gli effetti collaterali sono provocati da più specie di proteine che interagiscono con lo stesso farmaco.
La ricerca farmacologica tende ad ottenere farmaci il più possibile specifici, cioè che si leghino solo alla proteina che deve essere influenzata per modificare lo stato patologico verso la guarigione e quindi con la minore affinità possibile verso altre proteine al fine di ridurre al minimo gli effetti collaterali. I farmaci con minore numero di effetti collaterali (più specifici per la proteina bersaglio) in teoria sono quelli che interagiscono con la proteina nel sito di legame di un suo legante naturale (es. il sito di legame del substrato e i siti di regolazione da effettore). Questo perché il sito di legame del legante naturale è in genere come struttura e reattività, una parte unica di una o poche proteine ed esso è capace di riconoscere il proprio legante tra tutte le altre presenti nella cellula. Infatti il farmaco per interagire con la proteina nel sito di legame del substrato deve essere un analogo del substrato e per questo è meno probabile che interagisca casualmente nel sito catalitico o su altre parti della superficie di altre proteine cellulari. In genere i normali costituenti cellulari (nelle normali concentrazioni cellulari) interagiscono con sufficiente stabilità solo con le proteine che hanno siti specifici per legarli. Quindi i farmaci (e così altre molecole esogene) hanno maggiori probabilità di legarsi specificamente ad una proteina quanto più sono analoghi del legante naturale della proteina stessa. Tuttavia alcuni coenzimi (es. NAD, piridossalfosfato, ecc.) e substrati (es. ATP, glucosio-6P, ecc.) si legano fisiologicamente a più di una proteina e quindi anche i farmaci analoghi di substrati o coenzimi possono dare effetti collaterali. Il legame di un farmaco (e così di altre molecole esogene) può essere altamente specifico, cioè interessare solo una data proteina, anche quando esso interagisce con una zona della superficie della proteina e non su un sito di legame specifico per una funzione (esempio, sito catalitico). La

capacità della proteina di legare un prodotto di sintesi (non naturale) è casuale nel senso che non è geneticamente prevista dalla cellula. Essa non è stata selezionata in base alla sua capacità di legare o meno un composto che incontra per la prima volta; tuttavia l'interazione può essere altamente specifica ed essa è utilizzata dall'uomo per intervenire sull'attività della proteina.
Individui diversi possono avere gradi diversi di sensibilità al farmaco in relazione agli alleli posseduti, e quindi alla proteina codificata ed allo stato di omozigosi o di eterozigosi di un dato allele. Un individuo omozigote per l'allele che codifica la proteina con maggiore affinità per il farmaco sarà più sensibile al farmaco degli individui eterozigoti ed ancora di più degli individui omozigoti per l'allele che codifica una proteina con scarsa/nulla affinità per lo stesso farmaco.
Quanto descritto diviene più complesso quando uno stato patologico dipende dall'alterazione indipendente di più proteine che in condizioni normali contribuiscono ad una unica funzione. In questo caso si può ipotizzare che farmaci diversi (molecole diverse) possano contribuire alla riduzione della patologia avendo per bersagli proteine diverse responsabili della patologia. In questo caso, per essere certi di usare il farmaco giusto occorre individuare quale proteina deve essere influenzata. Con l'analisi genetica si possono individuare le varianti genetiche di una proteina e successivamente valutare la loro sensibilità ai farmaci e ad altre molecole. Tuttavia questo tipo di analisi non è sempre possibile perché spesso di una proteina non si conoscono tutte le varianti (genetiche o da modificazioni covalenti).
Lo studio del proteoma ha anche lo scopo di stabilire la diversa sensibilità ai farmaci, anestetici, allergeni, droghe delle proteine varianti codificate dai vari alleli della popolazione umana. Lo scopo è di arrivare ad usare farmaci ed anestetici in relazione alla genetica di ogni singolo paziente per conoscere preventivamente gli allergeni di ogni neonato e le possibilità di disintossicazione dalle droghe in relazione alla genetica di chi ne fa uso e di evitare decessi di pazienti intolleranti a farmaci o anestetici, innocui per la maggioranza della popolazione umana.
Il meccanismo, proposto per i farmaci sopra ed in figura D7), può essere esteso a tutte le altre molecole provenienti dall'ambiente volutamente od accidentalmente inalate, assunte per contatto od ingerite con gli alimenti. Queste molecole quando interagiscono con proteine possono attivarle, inibirle, denaturarle o favorirne la degradazione. Ed i loro effetti sono solo dannosi.

La classificazione delle molecole esogene in farmaci, veleni, droghe, ecc. non è in relazione al loro meccanismo di azione (interazione con componenti normali costituenti cellulari) ma in relazione all'effetto che esse provocano nell'organismo ed alle abitudini sociali della popolazione. Fumare tabacco è passato da piacevole abitudine a vizio. Successivamente è diventato un vizio pericoloso per il fumatore, poi anche per i non fumatori, ed è stato vietato nei luoghi pubblici.
Il proverbio giapponese: "Non tutto ciò che è saggio è piacevole", viene convertito in "Non tutto ciò che è piacevole è saggio" .

The most incomprensible thing about the world is that it is comprensible."

Albert Eistein, premio Nobel per la Fisica

Appendice E

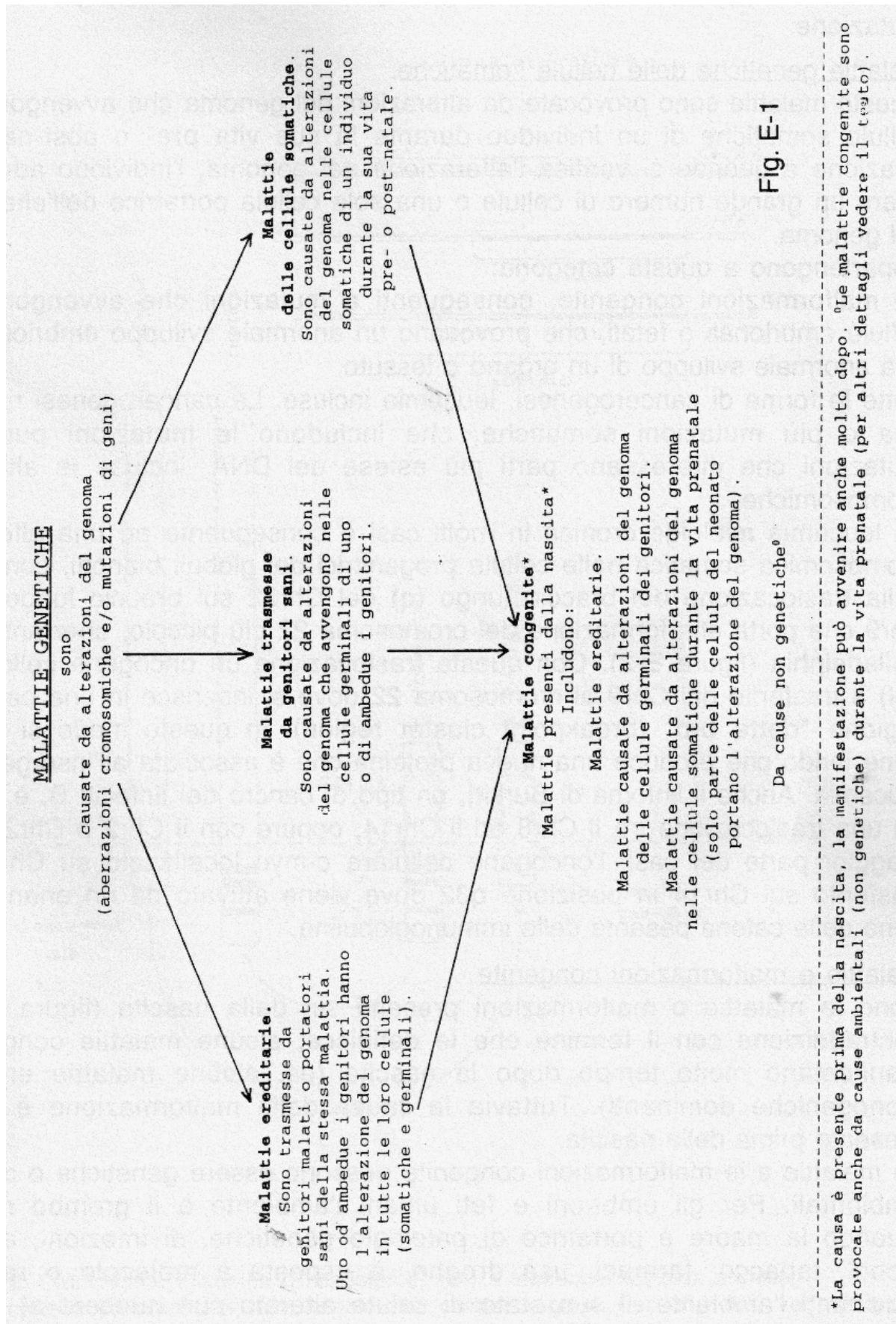
Malattie genetiche umane

Le malattie genetiche sono causate da una o più alterazioni del genoma (alterazioni dei cromosomi e mutazioni dei geni)(figura E-1). Esse includono:

Malattie ereditarie. I genitori, portatori sani o portatori della malattia, trasmettono ai figli l'alterazione del genoma responsabile della malattia. I genitori sono portatori sani se la malattia trasmessa è recessiva ed i genitori sono ambedue eterozigoti per il gene patologico. I genitori sono malati se sono omozigoti per un gene patologico recessivo oppure se sono portatori di un gene patologico dominante.

Malattie trasmesse da genitori sani che hanno subito un'alterazione del genoma solo nelle cellule germinali. I genitori non sono né malati né portatori sani della malattia, ma avendo subito un'alterazione del genoma proprio nelle loro cellule germinali causano l'insorgere della patologia nella loro prole. E' l'inizio di una malattia ereditaria. L'alterazione del genoma può avvenire casualmente durante la gametogenesi di un genitore, interessare uno o pochi gameti. Queste alterazioni del genoma si verificano per cause chimiche e fisiche, per errori (una base al posto di un'altra) durante la replicazione del DNA, o per alterazione della struttura o del numero dei cromosomi. Ad esempio può accadere che nella fase terminale della gametogenesi, durante la meiosi, si abbia la mancata segregazione di un cromosoma, per cui una cellula gametica risulta avere due copie dello stesso cromosoma. Questa cellula gametica darà luogo ad uno zigote trisomico per un dato cromosoma. Molti casi di trisomie portano ad aborti spontanei (specialmente quelle dei cromosomi più grandi). Nel caso della trisomia 21 (sindrome di Down) si ha la nascita di un individuo che ha una durata di vita quasi normale, molto ridotta quando sono presenti malformazioni cardiache od altre patologie. La sindrome di Down è causata nell'80% dei casi dalla madre e nel 20% dal padre. Il rischio della trisomia 21 aumenta considerevolmente tra i 35-45 anni di età della madre. La mancata segregazione del cromosoma 21 avviene quasi sempre alla prima divisione meiotica, ma può accadere anche alla seconda. Le patologie causate da alterazioni genetiche che avvengono durante la gametogenesi dei genitori sono congenite (presenti alla nascita) ma non ereditarie perché i genitori non sono portatori della patologia. Tuttavia esse possono essere l'inizio di una malattia familiare ereditaria se i figli malati possono procreare. Si può ipotizzare che il genoma delle cellule germinali di un individuo possa subire molto precocemente (durante la vita embrionale o fetale) una mutazione e che essa interessi tutti o gran parte dei precursori delle cellule germinali. Divenuto adulto, l'individuo avrà un soma normale, ma tutti o gran parte dei suoi gameti porteranno la mutazione.

Fig. E-1

*La causa è presente prima della nascita, la manifestazione può avvenire anche molto dopo. ⁰Le malattie congenite sono provocate anche da cause ambientali (non genetiche) durante la vita prenatale (per altri dettagli vedere il testo).

Malattie genetiche delle cellule somatiche.

Queste malattie sono provocate da alterazioni del genoma che avvengono nelle cellule somatiche di un individuo durante la sua vita pre- o post-natale. In relazione a quando si verifica l'alterazione del genoma, l'individuo adulto può avere un grande numero di cellule o una sola cellula portatrice dell'alterazione del genoma.

Appartengono a questa categoria:

- Le malformazioni congenite, conseguenti a mutazioni che avvengono nelle cellule embrionali o fetali, che provocano un anormale sviluppo embrionale e/o una anormale sviluppo di un organo o tessuto.
- Tutte le forme di cancerogenesi, leucemie incluse. La cancerogenesi risulta da una o più mutazioni somatiche, che includono le mutazioni puntiformi e le mutazioni che interessano parti più estese del DNA, incluse le alterazioni cromosomiche.

La leucemia mieloide cronica in molti casi è conseguente ad una alterazione cromosomica somatica nelle cellule progenitrici dei globuli bianchi, consistente nella traslocazione del braccio lungo (q) del Chr22 sul braccio lungo (q) del Chr9 che porta alla formazione del cromosoma 22 più piccolo, aberrante, detto Philadelphia (figura E-2). Con questa traslocazione un oncogene cellulare (c-abl) è trasferito dal Chr9 al cromosoma 22 dove si inserisce in una particolare regione "detta brc" (breakpoint cluster region). In questo modo si crea un gene ibrido che produce una nuova proteina che è associata all'insorgere della leucemia. Anche il linfoma di Burkitt, un tipo di cancro dei linfociti B, è causato da una traslocazione tra il Chr8 ed il Chr14, oppure con il Chr2 o Chr22. Nella maggior parte dei casi, l'oncogene cellulare c-myc localizzato su Chr8q24 è trasferito sul Chr14 in posizione q32 dove viene attivato da un enhancer del gene della catena pesante delle immunoglobuline.

Malattie e malformazioni congenite.

Sono le malattie o malformazioni presenti sin dalla nascita (figura E-1). In contraddizione con il termine che le definisce, alcune malattie congenite si manifestano molto tempo dopo la nascita (es. alcune malattie ereditarie, monogeniche dominanti). Tuttavia la causa della malformazione è sempre presente prima della nascita.

Le malattie e le malformazioni congenite possono essere genetiche o da cause ambientali. Per gli embrioni e feti umani l'ambiente è il grembo materno. Quando la madre è portatrice di patologie genetiche, di infezioni, abusa di alcool, tabacco, farmaci, usa droghe, è esposta a molecole o radiazioni inquinanti l'ambiente, il suo stato di salute alterato può nuocere al normale sviluppo dell'embrione e del feto (anche se geneticamente sano). Inoltre le stesse cause ambientali che danneggiano la madre possono influire direttamente sull'embrione o feto alterandone lo sviluppo.

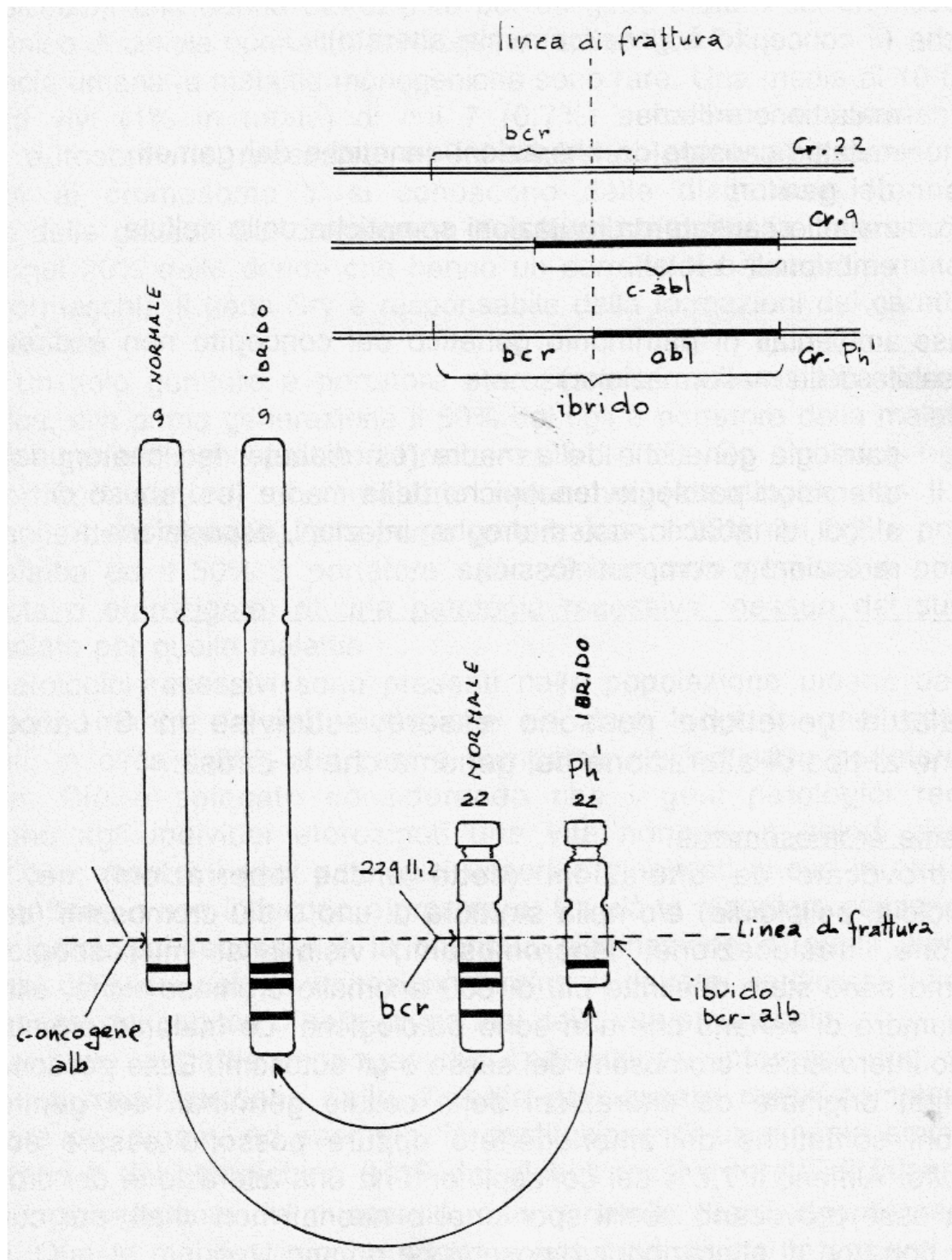



Figura E-2. Cromosoma Filadelfia (Philadelphia, Ph') e leucemia mieloide cronica.
c-abl, oncogene cellulare abl; bcr, breakpoint cluster region. Per altri dati vedere testo. (ridisegnato modificato da Connor J.M. and Fergusson-Smith,1991, Essential Medical Genetics, 3rd ed. Blackwell, London).

Sommario delle cause delle malattie e malformazioni congenite.

Genetiche (il concepito è geneticamente alterato):

-malattie ereditarie.
-malattie causate da alterazioni genetiche dei gameti dei genitori.
-malattie causate da mutazioni somatiche delle cellule embrionali o fetali.

Da cause ambientali (il patrimonio genetico del concepito non è direttamente responsabile della malformazione)

-patologie genetiche della madre (es. diabete, fenilchetonuria).
-alterazioni/patologie fenotipiche della madre (es. abuso di alcool, di tabacco, uso di droghe, infezioni, esposizione a radiazioni o composti tossici).

Le malattie genetiche possono essere suddivise in 3 categorie in relazione al tipo di alterazione del genoma che le causa:

1. Malattie cromosomiche.
Sono provocate da alterazioni (dette anche aberrazioni) nel numero (aneuploidia, poliploidia) e/o nella struttura di uno o più cromosomi (delezione, inversione, traslocazione, isocromosomi) visibili al microscopio ottico. Nell'uomo sono state descritte più di 600 anomalie cromosomiche, oltre ad un certo numero di varianti che non sono patologiche. Le malattie cromosomiche possono interessare i cromosomi del sesso o gli autosomi. Esse possono essere congenite, originate da alterazioni delle cellule germinali dei genitori o da mutazioni somatiche dell'embrione/feto oppure possono essere somatiche dell'adulto. Almeno il 7,5% dei concepimenti ha una alterazione dei cromosomi, tuttavia esse provocano aborti spontanei o neonati non vitali, per cui solo lo 0,6% dei portatori di alterazioni cromosomiche rimane vivo.

2. Malattie monogeniche .
Sono provocate da mutazioni di un singolo gene e possono essere: autosomiche dominanti, autosomiche recessive e legate al cromosoma X. Nelle donne, avendo esse due cromosomi X, le malattie legate a questo cromosoma si classificano come quelle autosomiche (dominanti o recessive), mentre nei maschi, i quali hanno un solo cromosoma X, la mutazione sarà sempre dominante. Il padre portatore di un cromosoma X mutato, lo trasmette a tutte le figlie (femmine) ed a nessun figlio (maschio). Poiché nelle donne uno dei due cromosomi X viene inattivato in tutte le cellule (esclusa la parte distale del

braccio piccolo), una donna eterozigote per un gene mutato sul cromosoma X è un mosaico di cellule normali e patologiche (appendice D).
Nella specie umana le malattie monogeniche sono rare. Una media di 10 casi su 1000 nati vivi (1% in totale) di cui 7 (0,7%) autosomiche dominanti, 2,5 (0,25%) autosomiche recessive e 0,4 (0,04%) legate al cromosoma X. Associate al cromosoma Y si conoscono delle disgenesie (formazione difettosa) delle gonadi. Sul cromosoma Y è localizzato Sry la cui mutazione è presente nel 20% delle donne che hanno un corredo XY (senza la mutazione sarebbero maschi). Il gene Sry è responsabile della formazioni dei caratteri del sesso maschile.
Quando un solo genitore è portatore eterozigote di una malattia dominante autosomica, alla prima generazione il 50% dei figli è portatore della malattia; se ambedue i genitori sono portatori l'incidenza sale al 75%. Se ambedue i genitori sono portatori sani di una malattia recessiva (eterozigoti per il gene responsabile della malattia), alla prima generazione il 25% dei figli è portatore della malattia ed il 50% è portatore sano. Se un solo genitore è portatore (omozigote o eterozigote) di una patologia recessiva, nessun dei suoi figli risulta malato per quella malattia.
I geni patologici recessivi sono presenti nella popolazione umana da molte generazioni, mentre la mutazione che ha dato luogo a geni patologici dominanti, in circa l' 80% dei casi è comparsa nell'individuo portatore della patologia. Ciò è spiegato considerando che i geni patologici recessivi permettono agli individui eterozigoti una vita normale e quindi anche la procreazione; mentre i geni patologici essendo manifesti anche in eterozigosi non permettono o non inducono a procreare. Da ciò la maggiore conservazione dei geni patologici recessivi e la continua eliminazione di quelli dominanti. Quindi nel 80% dei casi le mutazioni dominanti devono verificarsi durante la gametogenesi dei genitori o nelle prime fasi della vita embrionale.
Sebbene nelle malattie monogeniche l'alterazione interessi una singola proteina, la manifestazione della malattia può essere molto complessa ed interessare più organi. Ad esempio, la sostituzione di un singolo aminoacido nella catena β dell'emoglobina (HbS dei globuli rossi a forma di falce) causa anemia, ingrossamento del miocardio, danno renale, ittero, rigonfiamento dei linfonodi. Queste manifestazioni possono avere gradi diversi in pazienti diversi, e ciò è attribuito all'influenza di geni normali sulla patogenesi. Questa influenza è diversa perché ogni individuo è geneticamente diverso dagli altri.

3. <u>Malattie poligeniche e multifattoriali.</u>
Le malattie poligeniche dipendono dall'interazione del prodotto (proteina) di più geni posti su loci diversi ed indipendenti. Gli effetti delle proteine codificate dagli alleli di ciascuno di questi geni sono insufficienti a determinare la malattia, tuttavia gli effetti delle proteine codificate dai vari geni che concorrono ad instaurare la malattia si accumulano. Quindi gli alleli di ognuno di questi geni non sono di *per sé* patologici, ma sono patologiche alcune loro combinazioni. La malattia poligenica è detta <u>multifattoriale</u> quando l'insorgenza della malattia

dipende anche dall'interazione dei prodotti genici mutati con molecole (naturali o sintetiche) provenienti dall'ambiente, assunte con gli alimenti o somministrate come farmaci. Queste malattie sono definite genetiche perché ricorrenti nei membri di una stessa famiglia, tuttavia sono trasmesse in maniera complessa e non secondo le leggi di Mendel come le malattie monogeniche, proprio perché dipendono da più geni (ognuno dei quali segue le leggi di Mendel), dall'ambiente esterno e da quello interno (attività degli alleli dei geni non coinvolti nella patologia). La loro incidenza alla prima generazione varia tra il 5% ed il 10% (contro il 25% ed il 50% delle malattie monogeniche). Non conoscendo il numero dei geni coinvolti è difficile calcolare l'incidenza, essa varia da famiglia a famiglia ed in relazione diretta con la gravità della malattia ed il numero dei membri coinvolti. Si assume che un individuo riceva dai genitori una combinazione di alleli aventi pesi diversi nel determinare la malattia poligenica, se la somma di questi pesi supera un certo valore soglia si ha la manifestazione della malattia (presenza di sintomi clinici) che in relazione alla combinazione degli alleli può essere più o meno grave. Nella malattia multifattoriale, la soglia può essere superata per azione di uno o più fattori ambientali che contribuiscono all'insorgere della malattia. Le malattie multifattoriali sono molto eterogenee: il contributo relativo dei geni del rischio (della malattia) e dei fattori ambientali varia da paziente a paziente. Inoltre un singolo gene mutato può causare una malattie che in genere è multifattoriale. Ad esempio, le cardiopatie sono multifattoriali, tuttavia il 5% dei morti prematuramente per infarto sono eterozigoti per ipercolesterolemia familiare (monogenica) che produce aterosclerosi (alterazione delle arterie con infiltrazioni di colesterolo) in assenza di fattori ambientali straordinari che favoriscono le cardiopatie (inquinamento, fumo di sigarette, alcolici, stress, ecc.). Si assume che il peso di questo gene mutato sia maggiore di quello di altri geni mutati che cooperano alla patologia e da ciò la maggiore gravità della malattia e di incidenza di morte prematura.
Sono malattie multifattoriali: il cancro, l'ipertensione, la gotta, il diabete mellito (insulina dipendente ed insulina-indipendente), la schizofrenia, l'epilessia, la spina bifida, il labbro leporino ed altre ancora.

Alcune considerazioni sulla dominanza e sulla recessività delle malattie genetiche

Le malattie monogeniche dominanti hanno una espressione clinica variabile come severità, penetranza (frequenza di espressione del genotipo), e come organi interessati. Queste malattie possono comparire clinicamente solo nell'adulto. La penetranza dipende dall'età del portatore, come nel caso della malattia di Huntington. In alcuni casi (es. poliposi del colon), un individuo può avere il gene mutato ed un fenotipo normale per tutta la vita. Questa caratteristica è detta non penetranza della malattia ed è considerata il caso estremo di ritardo nel tempo della comparsa della malattia. La non penetranza

è un caso molto particolare, perché è un portatore sano che trasmette un carattere dominante (autosomico). Si dice che la malattia ha saltato una generazione.
In genere le malattie dominanti in eterozigosi sono meno severe delle malattie recessive.
Anche le malattie recessive hanno espressione clinica e penetranza variabili, ma in genere esse hanno caratteristiche cliniche più uniformi e si instaurano più precocemente di quelle dominanti. Le malattie recessive si manifestano e sono diagnosticate prevalentemente nei bambini mentre quelle dominanti negli adulti.
Le malattie autosomiche dominanti spesso risultano da mutazioni che interessano proteine recettoriali, strutturali ed enzimi regolati a inibizione retrogada o appartenenti a vie sintetiche. Le malattie autosomiche recessive spesso risultano da mutazioni che interessano enzimi non regolati di vie cataboliche.
Gli individui portatori omozigoti del gene mutato, sono malati perché i due alleli mutati non producono la proteina o producono quantità di proteina attiva insufficienti a rispondere alle normali esigenze dell'organismo. Esiste un secondo tipo di allele mutato, che rispetto all'allele normale produce una quantità minore di proteina attiva, ma che in omozigosi od in eterozigosi con l'allele normale, produce una quantità sufficiente a rispondere alle normali esigenze dell'organismo. Tuttavia quando questo allele forma un composto genetico (omozigosi di alleli mutati in modo/posizioni diverse) con un allele mutato (il cui prodotto genico è scarso o nullo) la proteina attiva prodotta risulta insufficiente per cui si ha l'insorgere della malattia. Con questa spiegazione si assume che ogni specie di proteina abbia un valore soglia di attività totale per cellula al di sotto del quale si ha l'insorgenza della malattia ed al di sopra del quale esiste una gamma di valori che conferiscono lo stato normale. In relazione al tipo di proteina mutata, la patogenesi instaurata da questi composti genetici può essere spiegata anche con altri meccanismi (vedere dopo emoglobina patologica HbSC)
In alcuni casi, la gamma di valori normali può avere anche una soglia superiore che se superata può dare un'altra manifestazione patologica (es. oncogeni).
La gamma dei valori normali di attività biologica di una proteina può essere più o meno estesa in relazione della funzione svolta. Ad esempio, le attività catalitiche degli enzimi glicolitici per i muscoli in condizioni di riposo sono molto basse se paragonate a quelle che hanno gli stessi enzimi in condizioni di un intenso sforzo muscolare. Questo può spiegare come in alcuni individui uno stato patologico si manifesti in alcune condizioni fisiologiche dell'organismo (es. sforzo muscolare, digiuno, ecc.) e non in altre (riposo muscolare, normale alimentazione, ecc.).
Si assume che le mutazioni recessive in genere interessino enzimi di vie cataboliche mentre le mutazioni dominanti interessino proteine non enzimatiche od enzimi regolati con l'inibizione retrogada. La mutazione recessiva può essere spiegata considerando che la perdita di attività catalitica causata dalla mutazione

di un solo allele possa essere compensata con l’attivazione dei normali meccanismi di regolazione da substrato e da inibizione retrogada che regolano le vie metaboliche e compensano le variazioni di attività catalitica di enzimi causate dalla variabilità genetica di alleli non patologici (figura D-6)
Se assumiamo che la mutazione recessiva spenga completamente l'espressione dell'allele o causi la perdita completa dell'attività enzimatica della proteina da esso codificata, in condizioni di eterozigosi, l'attività totale per cellula dell'enzima sarà solo il 50% di quella normale, perché prodotta dall'enzima espresso dall'allele normale. Questa attività potrà essere incrementata mediante la regolazione da substrato (per maggiore saturazione dell'enzima) e/o per disinibizione dell'enzima che regola a inibizione retrogada la via metabolica. Per qualsiasi motivo il prodotto finale di una via metabolica sia carente per le esigenze della cellula e/o dell'organismo, il meccanismo a inibizione retrogada fa incrementare il flusso metabolico di quella via. A queste compensazioni va aggiunto l'incremento di stabilità e quindi di attività dell'enzima conseguente la sua maggiore saturazione da substrato.
Mutazioni che interessano il primo od il secondo enzima di vie cataboliche che degradano i metaboliti provenienti per via ematica dagli alimenti, come la fenilalanina ed il galattosio causano malattie recessive piuttosto che dominanti.
La fenilchetonuria, malattia causata dalla mutazione dell'enzima fenilalanina idrossilasi (figura E-4), è un esempio di compensazione da substrato della perdita del 50% dell'attività enzimatica. L'enzima fenilalanina idrossilasi è il primo enzima della via di degradazione della fenilalanina, ed è regolato da substrato e mediante fosforilazione. In condizione di eterozigosi, la ridotta attività dell'enzima provoca un incremento di circa due volte della concentrazione allo stato stazionario della fenilalanina nel sangue e nelle cellule. Ciò porta ad una maggiore saturazione dell'enzima fenilalanina idrossilasi, che diviene più attivo e quindi compensa la perdita del 50% di enzima provocata dalla mutazione. Pertanto non si ha la comparsa di sintomi clinici. L'incremento della CSS della fenilalanina nel sangue e nelle cellule è in relazione inversa all'attività dell'enzima, per cui si arresta quando la velocità del catabolismo della fenilalanina eguaglia il flusso di sangue nella cellula epatica. In condizioni di omozigosi degli alleli mutati, la compensazione non è più possibile perché si ha una forte riduzione/assenza dell'attività enzimatica che non può essere compensata neanche da un enorme incremento della concentrazione di fenilalanina. La fenilalanina e suoi metaboliti (es. acido fenilpiruvico) rimangono in alta concentrazione nel sangue e nelle cellule e finiscono per causare danni irreparabili al sistema nervoso del neonato perché l'alta concentrazione causa l'associazione a molecole e strutture sopramolecolari non possibile alle concentrazioni fisiologiche. La malattia è provocata dal fallimento del meccanismo di compensazione (data la scarsa attività residua dell'enzima) che agisce proprio incrementando la concentrazione di fenilalanina, e poi dalla tossicità dell'aminoacido e del fenilpiruvato. Se la loro tossicità fosse stata più alta ed avesse provocato gli stessi danni a concentrazioni più basse (come quelle osservate in eterozigosi), la fenilchetonuria sarebbe stata una malattia

dominante. Se la fenilalanina e i suoi metaboliti non fossero tossici e/o fossero eliminati normalmente con le urine non ci sarebbe questa patologia.
Gli enzimi regolati a inibizione retrogada hanno alcune caratteristiche che rendono più difficile la loro compensazione: sono i pace-maker e rate limiting step (decidono la velocità e sono limite massimo della velocità) della via metabolica, pertanto hanno minore concentrazione e maggiore turnover degli enzimi non regolati, maggiore dipendenza dalla regolazione da effettore piuttosto che da quella da substrato. La limitazione appare maggiore per gli enzimi regolati a inibizione retrogada appartenenti a vie metaboliche sintetiche. Questo probabilmente perché le vie sintetiche sono più complesse di quelle di degradazione e richiedono energia sotto forma di ATP od altri nucleotidi trifosfati e potere riducente in genere fornito da NADH o NADPH. Inoltre questi metaboliti e coenzimi, essendo coinvolti in molte vie metaboliche sintetiche, hanno concentrazioni cellulari mantenute entro limiti ben precisi. In particolare la CSS del ATP è un importante segnale dello stato energetico della cellula e l'ATP è effettore allosterico di molti enzimi glicolitici. Il rapporto delle CSS NADH/NAD è segnale dello stato redox della cellula.
Da tutto ciò si ha l'indicazione delle possibili cause della dominanza delle mutazioni che interessano alcuni enzimi regolati a inibizione retrogada di vie sintetiche. La malattia può essere causata dalla perdita del prodotto finale o dall'accumulo di intermedi dannosi alla cellula o all'organismo.
Un esempio di una patologia causata dalla mutazione dominante di enzima rate limiting step coinvolto in una sintesi è data dalla porfiria acuta intermittente, causata dalla carenza di porfobilinogeno-deaminasi, enzima della via di sintesi dell'eme nel fegato. Sebbene in eterozigosi, cioè con una attività catalitica residua (teorica) uguale ad almeno il 50% del normale, essa risulta insufficiente per le normali attività metaboliche perché non compensata. Si ipotizza che la mutazione risulti dominante quando essa altera proprio la proprietà di regolazione dell'enzima. Ad esempio, se la mutazione rende l'enzima regolato più affine all'effettore negativo (il prodotto finale della via metabolica), l'enzima mutato rimane cronicamente inibito anche in condizioni in cui il prodotto finale è scarso. Per compensare la perdita di attività catalitica può essere disinibito solo l'enzima prodotto dall'allele normale, tuttavia la concentrazione degli enzimi regolati, essendo programmata per essere il passaggio che limita la velocità della via metabolica, non è in eccesso come l'attività catalitica degli enzimi non regolati. Quindi probabilmente per certi enzimi regolati essa non può essere incrementata fino al punto da poter compensare la perdita del 50% dell'attività catalitica totale.
Una mutazione che renda un enzima regolato a inibizione retrogada meno affine al PF può risultare in un eccesso di PF che non può essere compensato e che può provocare alterazioni della fisiologia cellulare. PF pur essendo un normale costituente cellulare, se presente in eccesso, può associarsi ad una o più proteine cellulari ed influenzare la loro attività biologica causando effetti negativi alla cellula in cui è prodotto; inoltre PF può uscire dalla cellula che lo produce e riversarsi nel sangue per poi influenzare negativamente la fisiologia

di altre cellule (come osservato per la fenilchetonuria). Altri tipi di proteine che, se mutate, sono responsabili anche in eterozigosi di patologie dominanti sono i recettori di membrana e le proteine di trasporto.
Un esempio di malattia dominante causata da una mutazione su una proteina non enzimatica è l'ipercolesterolemia familiare. La mutazione interessa il recettore per le lipoproteine a bassa densità (LDL) e provoca l'aumento del livello ematico delle LDL con conseguente deposito di colesterolo nelle arterie (causando una prematura ischemia cardiaca) e nei tendini. I recettori sono presenti sulla membrana cellulare in numero geneticamente definito e le cause della dominanza appaiono nella perdita (non compensabile) di metà dei recettori. Il ridotto numero di recettori non permette la normale entrata del colesterolo nelle cellule di vario tipo che lo utilizzano per alcune sintesi (ormoni steroidei, vitamina D, acidi biliari). Le sintesi possono egualmente avvenire ma cambia il livello ematico del colesterolo. Quindi a causa del cronico alto livello ematico di un normale costituente cellulare (colesterolo) si instaura una patologia.
Le cause che determinano la maggiore penetranza e la maggiore costanza di manifestazione clinica di una malattia monogenica recessiva in individui omozigoti rispetto a quelle dominanti in individui eterozigoti non sono chiare. Una spiegazione può essere data assumendo che in omozigoti le malattie recessive risultino da una perdita di attività biologica quantitativamente maggiore rispetto a quella causata da una malattia dominante in individui eterozigoti. In questa ipotesi, gli effetti delle mutazioni recessive in omozigoti sarebbero più difficilmente compensabili di quelli causati da un singolo allele patologico dominante. Le malattie recessive creerebbero scompensi lontani dalla capacità di compensazione dei meccanismi cellulari, per cui si manifesterebbero precocemente e con caratteristiche più costanti. Le malattie dominanti in individui eterozigoti subirebbero compensazioni quantitativamente maggiori anche in relazione all'eterogeneità genetica degli individui portatori dell'allele mutato ed alla loro età. Da ciò la diversa penetranza, severità ed espressione clinica di una stessa malattia dominante in individui diversi.

Eterogeneità genetica di una stessa malattia

Si ritiene che la maggior parte delle malattie genetiche umane siano geneticamente eterogenee.
L'eterogeneità genetica di una malattia (stesse manifestazioni cliniche o molto simili) può essere:
1. Eterogeneità genica (da geni diversi). La malattia è prodotta indipendentemente da geni diversi mutati (mutazioni su loci diversi). Uno solo di questi geni causa la malattia (se occorrono più geni mutati per causarla, la malattia è classificata come multifattoriale).
2. Eterogeneità allelica (da alleli diversi). La malattia è prodotta sempre dallo stesso gene mutato (unico locus), tuttavia del gene esistono più alleli con

mutazioni diverse come posizione o base sostituita nella stessa posizione (poliallelismo patologico). Alleli diversi esprimono la stessa proteina mutata ma con aminoacidi sostituiti in posizioni diverse, oppure nella stessa posizione sono stati sostituiti aminoacidi diversi.
3. Composti genetici. Un caso particolare dell'eterogeneità allelica è il composto genetico, si ha omozigosi dello stesso gene mutato, ma i due alleli hanno mutazioni diverse. In genere la malattia causata dal composto genetico ha caratteristiche intermedie tra quelle prodotte dagli omozigoti dei due alleli mutati; talvolta invece ha caratteristiche più vicine a quelle della malattia causata da uno dei due alleli mutati.

Alcuni esempi di malattie con eterogeneità genetica

Un esempio di malattia con eterogeneità genica, provocata da mutazioni su geni diversi (loci diversi), è dato dall'emofilia, difetto della coagulazione del sangue, della quale sono responsabili indipendentemente due differenti geni mutati, ambedue sul cromosoma X. Uno è responsabile della deficienza del fattore VIII (ChrXq28)(emofilia classica), l'altro del fattore IX (ChrXq27)(emofilia di Christmas)(figura E-3).
Altri casi di eterogeneità genica sono la fenilchetonuria, la malattia può essere causata indipendentemente da 3 geni mutati (figura E-4) e la patologia MODY2 da almeno sei geni diversi (capitolo 4).
La Met-emoglobinemia ereditaria è caratterizzata da eterogeneità genica ed allelica (10 mutazioni indipendenti su 3 differenti loci genetici). L'emoglobina è un tetrametro costituito da due subunità α e due β. Questa patologia può essere causata da emoglobine mutate (dette emoglobine M) o da un enzima mutato: citocromo-b5 riduttasi, chiamato anche anche Met-Hb riduttasi (figura E-7). Esistono almeno cinque emoglobine M, originate da mutazioni indipendenti: due sulla subunità α (Chr16p), tre sulla subunità β (Chr11p) dell'emoglobina. Ed almeno 5 nel locus (Chr22) del gene che codifica l'enzima Met-Hb riduttasi.
Un esempio di composto genetico è la emoglobinopatia SC. S sta per emoglobina (Hb) dell'anemia a cellule falciformi il cui glutammato in posizione 6 della subunità β (β-6Glu) è sostituito con una valina. La forma a falce è causata dal fatto che l'Hb-S deossigenata tende a polimerizzare formando dei filamenti che deformano i globuli rossi, i quali da rotondi prendono la forma a falce. L'HbC è l'emoglobina in cui il β-6Glu è sostituito con una lisina. L'HbC dà sintomi clinici minori e diversi rispetto a quelli dell'Hb-S, tuttavia la subunità β-C (e non la subunità β) copolimerizza con la subunità β–S, per cui il composto genetico (β–S β-C) mostra le stesse alterazioni dell'omozigote β-S β-S, anche se in forma un poco più attenuata. L'emoglobinopatia SC è autosomica recessiva per cui si manifesta in condizioni di omozigosi, tuttavia quando l'allele normale β è sostituito con un β-C la malattia si manifesta egualmente. L'allele β–C in omozigosi non è falcemizzante, ma lo é quando l'allele β–S é sul cromosoma omologo.

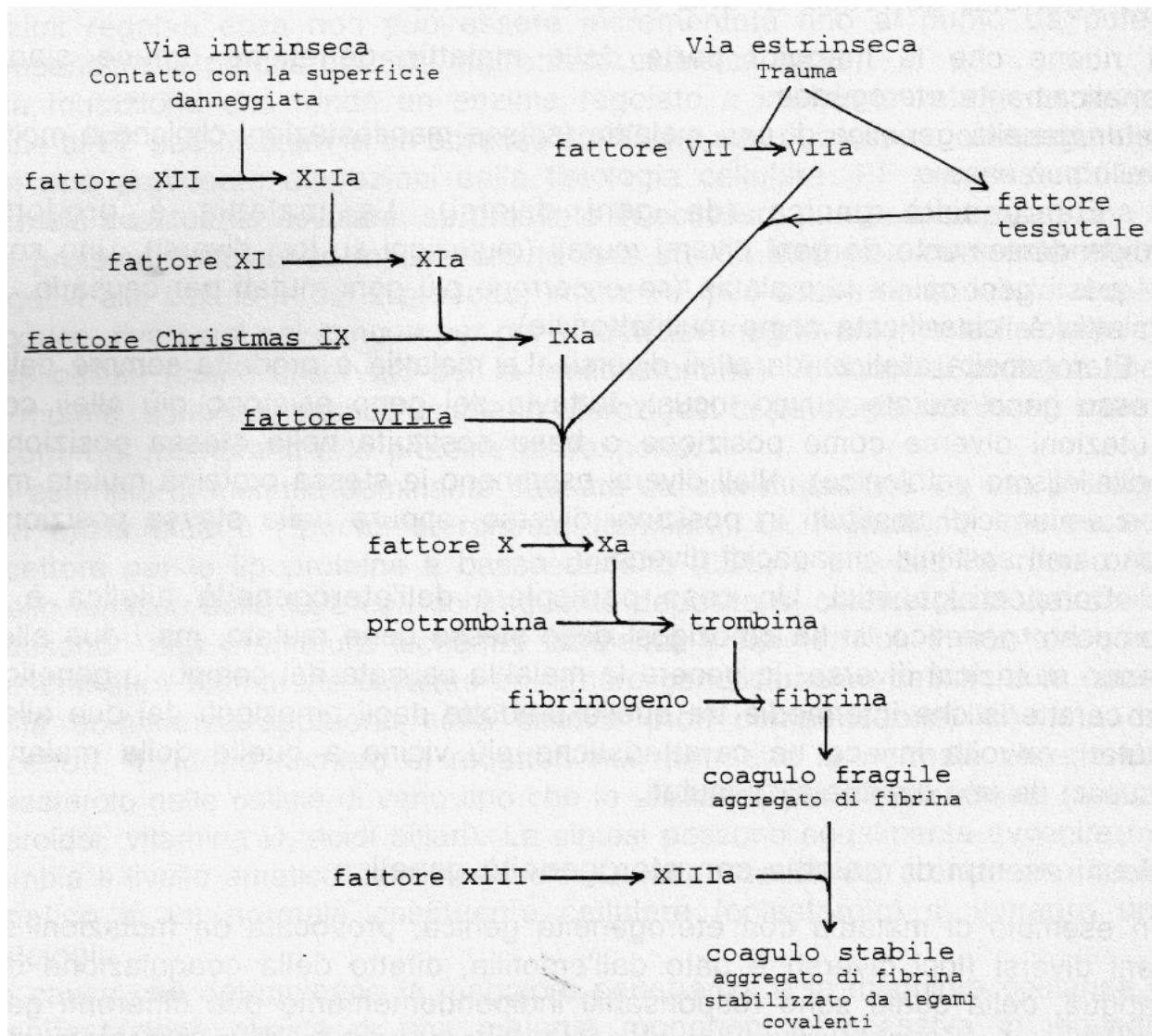



Figura E-3.
Schema del meccanismo molecolare di formazione del coagulo. Lo schema trascura le reazioni di attivazione a feedback ed altre non ancora definitivamente provate. I fattori di coagulazione sono proteasi, eccetto il fattore VIII che non ha attività enzimatica, ed il fattore XIII che catalizza la formazione di legami covalenti tra le molecole di fibrina del coagulo fragile. La forma attiva (indicata con a) di un fattore attiva il fattore successivo. Il meccanismo è detto a cascata e provoca rapidamente l'amplificazione del numero dei prodotti di reazione ad ogni passaggio. La via intrinseca è attivata da una superficie irregolare dei tessuti, quella estrinseca da sostanze liberate dai tessuti traumatizzati. I due fattori responsabili dell'emofilia sono sottolineati (da Gabrielli F., Querci B. e Bolognani L. (1983), Modificazioni post-traduzionali delle proteine Piccin, Padova. Ridisegnato e modificato).

Alcuni esempi di malattie genetiche umane

La fenilchetonuria (PKU) è una malattia monogenica autosomica recessiva con eterogeneità genica su tre loci (figura E-4). E' causata indipendentemente dalla mutazione di tre enzimi epatici diversi: 1) fenilalanina idrossilasi (locus Chr12q1), 2) biopteridina sintetasi e 3) diidropteridina riduttasi. Il difetto funzionale che risulta in ognuno dei tre casi è l'incapacità del fegato a metabolizzare la fenilalanina. La conversione della fenilalanina in tirosina si realizza poco o non si realizza in conseguenza della minore o nulla attività dell'enzima-1 (fenilalanina idrossilasi). In altri individui l'enzima 1) è normale nella attività molecolare e nella concentrazione, ma non può funzionare perché il suo coenzima biopteridina non è sintetizzato a sufficienza (é mutato l'enzima-2), oppure in altri individui la biopteridina è sintetizzata normalmente ma non viene ridotta a sufficienza (è mutato l'enzima-3). Il blocco della via metabolica determina un incremento abnorme nel sangue di fenilalanina e di altri metaboliti (es. fenilpiruvato) che normalmente sono presenti in concentrazioni più basse, ma che ad alte concentrazioni causano danni irreversibili alle cellule nervose dei neonati. Il fegato non è danneggiato dal blocco della sua via metabolica, perché la tirosina per la sintesi proteica epatica è assunta con la dieta mentre l'ATP, non prodotto dalla mancata ossidazione della fenilalanina, non crea carenze perché può essere sintetizzato a sufficienza per le normali vie per la sintesi di ATP (glicolisi e fosforilazione ossidativa). Infatti la funzione della via metabolica è proprio l'eliminazione ossidativa degli eccessi di fenilalanina e tirosina, data la loro tossicità, e non la sintesi di un prodotto finale specifico della via metabolica (cioè prodotto solo da essa) non sostituibile con altri metaboliti e necessario per la fisiologia cellulare e/o dell'organismo. L'ATP è un prodotto non sostituibile e necessario alla cellula, tuttavia è prodotto soprattutto da altri metaboliti. La quasi totalità del ATP è prodotto dall'ossidazione del glucosio e dei grassi, mentre il contributo alla sintesi di ATP dato dall'ossidazione degli aminoacidi è scarso (eccetto che nelle diete iperglucidiche) ed è particolarmente scarso il contributo della fenilalanina, aminoacido presente in piccole quantità nelle proteine.

Sono state osservate mutazioni di geni codificanti altri enzimi della stessa via metabolica:
tirosina amino-transferasi (4 in figura E-4) se insufficiente causa il blocco/riduzione del metabolismo della tirosina con conseguente incremento della tirosina nel sangue (tirosinemia) e di altri metaboliti. La malattia è chiamata tirosinemia ed è caratterizzata da ipercheratinizzazione, lesioni degli occhi e della pelle e neuropatie;
omogentisico ossidasi (5 in figura E-4) se insufficiente causa l'accumulo dell'acido omogentisico, che è convertito in un composto polimerico nero che ha affinità per i tessuti contenenti collagene. La malattia è chiamata alcaptonuria per il colore nero delle urine dei pazienti, essa provoca anche la pigmentazione nera, l'alterazione del tessuto connettivo ed artrite. (prima

correlazione tra una patologia per carenza di un enzima e una alterazione genetica,

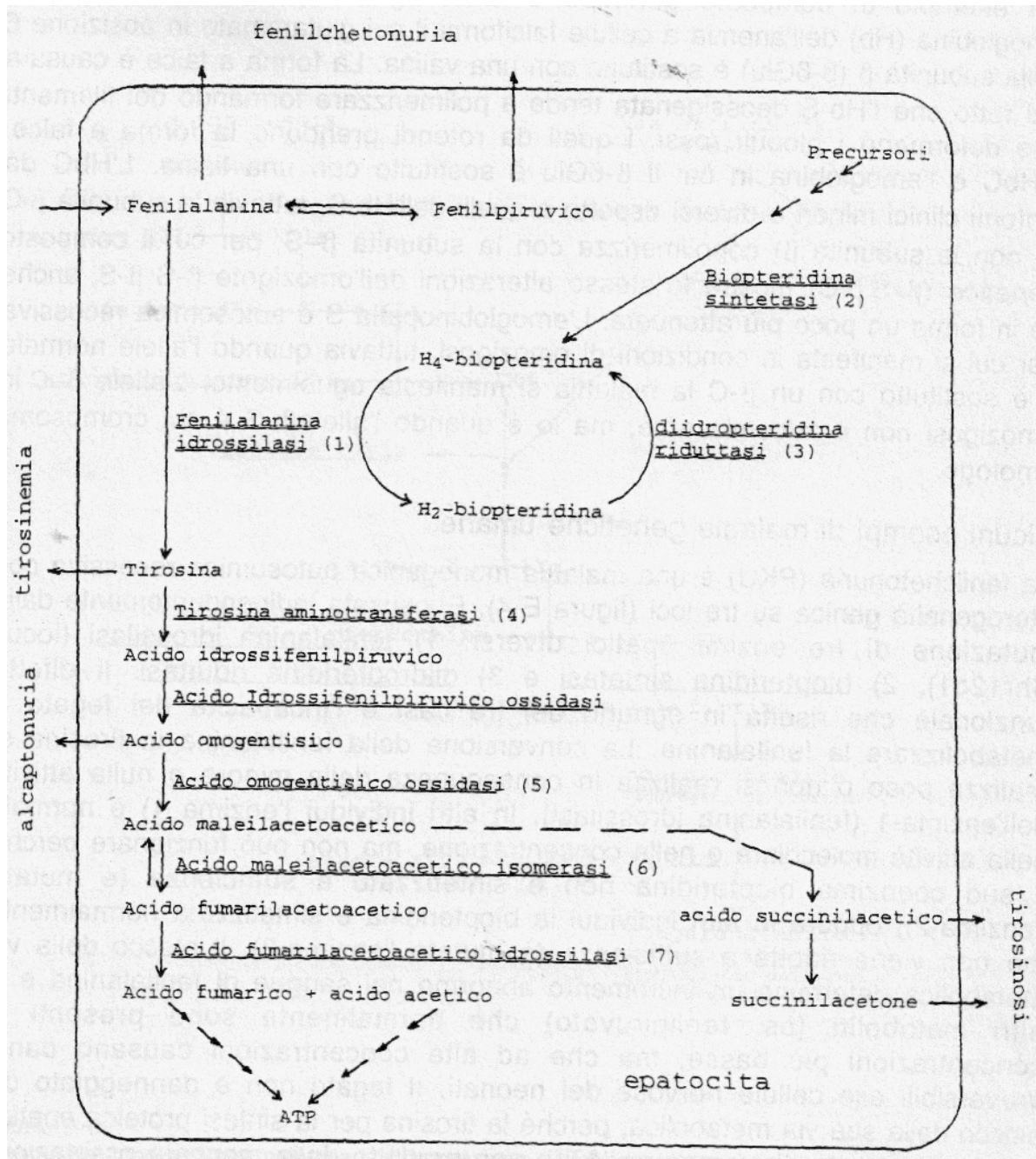



Figura E-4. Alterazioni del metabolismo ossidativo della fenilalanina nel fegato umano. Le frecce che escono dall'epatocita indicano i metaboliti che fuoriescono dalle cellule nei vari blocchi metabolici. Per altri dati vedere il testo.

fatta nel 1908 da Sir Archibald Garrod che per queste malattie coniò il termine "Errori congeniti del metabolismo";
maleilacetoacetico isomerasi (6 in figura E-4) e fumarilacetoacetico idrossilasi (7 in figura E-4) se insufficienti causano l'incremento della concentrazione dei loro substrati e quindi la formazione di acido succinilacetoacetico e succinilacetone. Questi composti causano l'alterazione della fisiologia dei tubuli renali e del fegato e neuropatie. La malattia è chiamata tirosinosi.
Anche in queste quattro alterazioni metaboliche, come nella fenilchetonuria, i sintomi clinici provengono da alterazioni provocate dall'accumulo di intermedi e non dalla minore sintesi di ATP.
Nei melanociti, la tirosina attraverso una specifica via metabolica è convertita in prodotti finali (melanine) utili all'organismo come decorazione e come protezione contro le radiazioni del sole (figura E-5). L'alterazione che rende completamente inattivo l'enzima tirosina idrossilasi, detto anche tirosinasi, causa l'albinismo di tipo I: completa mancanza di pigmento nella pelle, occhi e capelli, in presenza di una normale distribuzione ed un normale numero di melanociti.
Nel tessuto nervoso (figura E-6), la tirosina viene convertita in neurotrasmettitori (adrenalina e noradrenalina) ed anche in neuromelanina (pigmento del sistema nervoso). Il primo enzima di questa via è l'enzima tirosina idrossilasi, tuttavia esso funziona anche nei soggetti albini suggerendo che le due vie metaboliche (per la sintesi delle melanine e dell'adrenalina, noradrenalina e neuromelanina) utilizzino enzimi tirosina idrossilasi codificati da geni diversi. Di recente è stata dimostrata l'alterazione genetica dell'enzima tirosina idrossilasi dei neuroni dopaminergici nigrostriati che causa "distonia sensibile alla DOPA" (alterazione del tono muscolare che è ridotta da somministrazioni di DOPA). La stessa sintomatologia risulta dall'alterazione congenita dell'enzima GTP-cicloidrolasi (enzima che fa parte della via metabolica di sintesi della H_4-biopteridina cofattore della tirosina idrossilasi) e tirosina idrossilasi dei neuroni dopaminergici nigrostriati che causa "distonia sensibile alla DOPA" (alterazione del tono muscolare che è ridotta da somministrazioni di DOPA). La stessa sintomatologia risulta dall'alterazione congenita dell'enzima GTP-cicloidrolasi (enzima che fa parte della via metabolica di sintesi della H_4-biopteridina cofattore della tirosina idrossilasi) e dalla mancanza congenita dei neuroni nigrostriati. Quindi abbiamo tre alterazioni genetiche diverse che causano la stessa sintomatologia.
La Met-emoglobinemia è una malattia genetica recessiva causata dalle emoglobine M (Hb-M). Queste emoglobine hanno mutazioni su una delle due subunità (α o β) che rendono l'ossidazione spontanea del ferro eminico (da Fe^{2+} a Fe^{3+}) più frequente rispetto alle emoglobine normali (figura E-7a). Le mutazioni interessano sia la catena α che la β, cioè due geni diversi (eterogeneità genica), inoltre ambedue le catene possono essere mutate in posizioni diverse (eterogeneità allelica) nell'istidina prossimale F8 o in quella

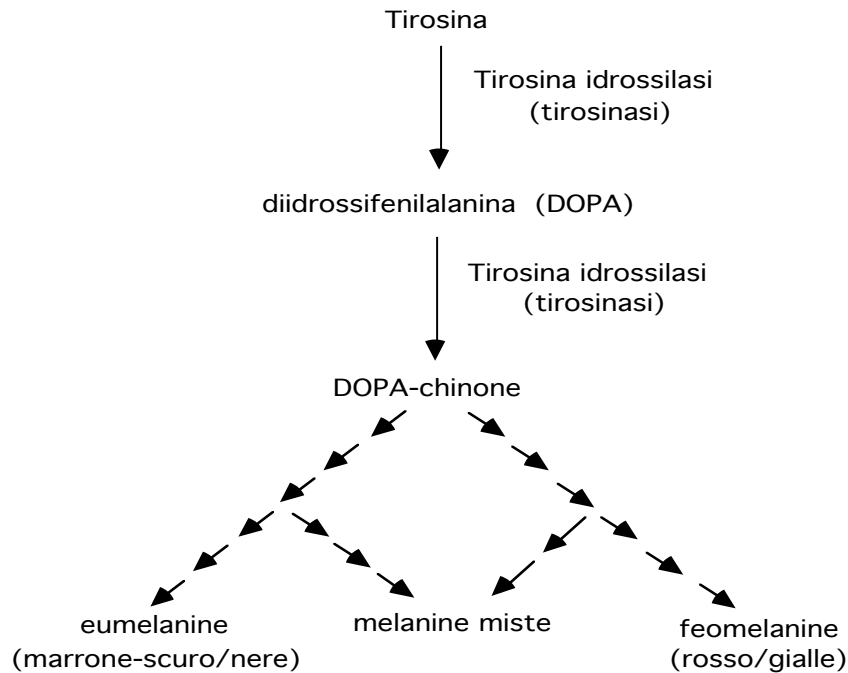



Figura E-5.
Schema della sintesi delle melanine nei melanociti (pelle, occhi, capelli, peli). La mancanza totale della tirosina idrossilasi causa l'albinismo di tipo I (mancanza totale di pigmento nella pelle, capelli, ecc.). La parziale attività dello stesso enzima provoca albinismo di tipo II (produzione di piccole quantità di pigmento che aumenta con l'età). Negli albini di colore "giallo" si ha una maggiore sintesi di feomelanine e mancano le eumelanine.

---

distale E7 (l'istidina, il cui residuo interagisce con il Fe eminico, è sostituita con la tirosina). Una forma di emoglobina M è data dalla sostituzione della valina β-E11 (il cui residuo idrofobico è posto vicino al ferro) con glutammato. Queste 5 alterazioni della struttura delle Hb-M incrementano la velocità di ossidazione spontanea del ferro eminico in modo tale che il normale sistema di riduzione dell'emoglobina ossidata dei globuli rossi non è più sufficiente per ridurre tutte le molecole ossidate . La Met-emoglobinemia si verifica anche per mutazioni dell'enzima citocromo-b5 riduttasi (figura E-7b) in portatori di Hb normali. Questo enzima ha la funzione di ridurre l'emoglobina normale che occasionalmente ma inevitabilmente si ossida nello svolgere la sua funzione di trasportatore di O_2. Infatti l'O_2 lasciando il ferro eminico (dell'Hb) a cui è legato, talvolta sottrae un elettrone al ferro, divenendo ione superossido (O_2^-). Il ferro eminico passando dallo stato di ossidazione due (Hb-F^{2+}) a tre (Hb-F^{3+}) rende l'emoglobina incapace ad associare di nuovo l'O_2. Pertanto l'emoglobina ossidata (Met-Hb) è incapace a svolgere la sua funzione di trasportatore di O_2 ed ha un colore diverso dalla Hb normale. I portatori di questa patologia hanno la colorazione bluastra della pelle e delle mucose (cianosi). Negli individui normali,

la velocità di ossidazione spontanea della Hb è 250 volte inferiore a quella della reazione di riduzione catalizzata dall'enzima NAD citocromo-b5 riduttasi per cui, nel sangue, il contenuto di Met-Hb è solo l'1% dell'Hb totale. Se l'enzima Met-Hb riduttasi è mutato la Met-Hb sale al 25-50%.
Quindi lo stesso stato patologico (Met-emoglobinemia) può essere causato da mutazioni dell'enzima citocromo-b5 riduttasi che deve riparare le Hb che si ossidano spontaneamente, oppure da Hb mutate (Hb-M), che hanno un'alta velocità di ossidazione. In ambedue i casi la patologia è recessiva.
L'enzima citocromo-b5 riduttasi normale ha attività sufficiente per ridurre la quantità di Hb normale che si ossida spontaneamente in condizioni normali, mentre non è sufficiente a ridurre le Hb-M il cui il ferro eminico si ossida a più alta velocità e non è sufficiente a ridurre l'Hb normale quando è resa più facilmente ossidabile da composti esogeni che interagiscono direttamente con l'Hb stessa (vedere dopo fenocopie di malattie genetiche).

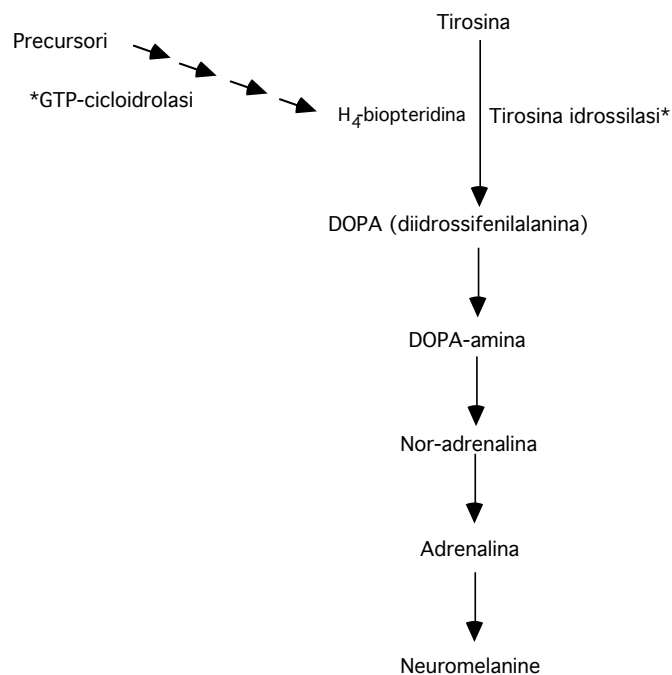



Figura E-6. Schema della sintesi dei neurotrasmettitori (adrenalina e noradrenalina) e della neuromelanina (pigmento del sistema nervoso). Nei neuroni nigrostriati dopaminergici dei gangli basali l'alterazione genetica dell'enzima tirosina idrossilasi o dell'enzima GTP-cicloidrolasi (questo ultimo fa parte della via metabolica di sintesi della H_4-biopteridina cofattore dell'enzima tirosina idrossilasi) è responsabile della patologia "distonia sensibile alla DOPA" Questo è un esempio di eterogeneità genica di una patologia.

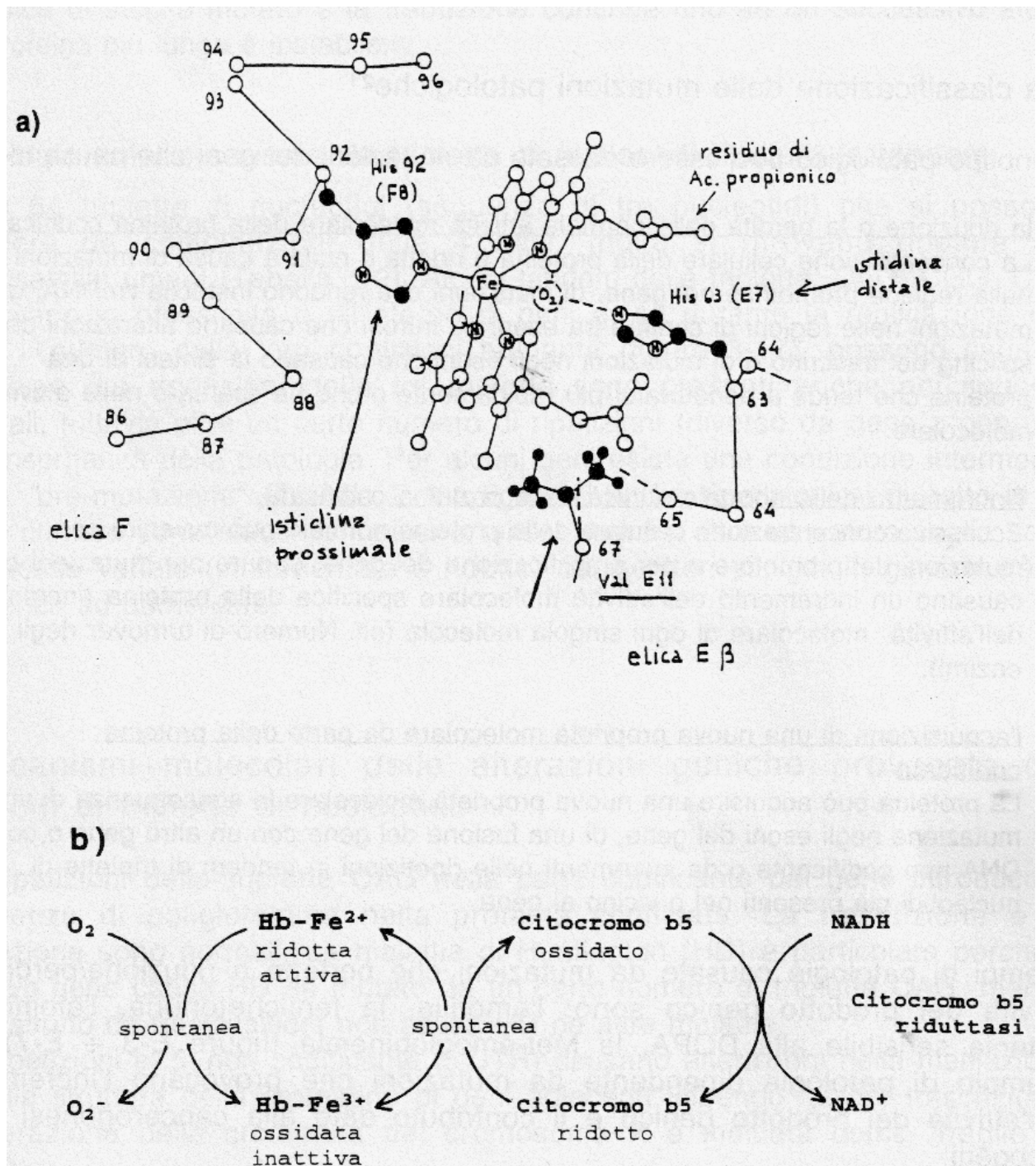



Figura E-7.
a) Regione dell'eme di una subunità β del Hb. La freccia indica il residuo di valina che è sostituito da uno di glutammato in una forma di Hb-M. I residui di istidina sono sostituiti indipendentemente da residui di tirosina in alcune forme di Hb-M.
b) Ossidazione spontanea dell'emoglobina normale e sua rigenerazione operata dall'enzima citocromo-b5 riduttasi.

Una classificazione delle mutazioni patologiche

Il fenotipo patologico può essere causato da mutazioni dei geni che causano:

la riduzione o la perdita della normale attività molecolare della proteina codificata.
La concentrazione cellulare della proteina è ridotta o nulla a causa di mutazioni nella regione promotrice del gene, di mutazioni che rendono instabile l'mRNA, di mutazioni nelle regioni di confine tra esoni ed introni che causano alterazioni dello splicing del trascritto o di mutazioni negli esoni che causano la sintesi di una proteina che tende a denaturarsi più rapidamente o che ha scarsa o nulla attività molecolare.

l'incremento della normale attività della proteina codificata.
Una eccessiva concentrazione cellulare della proteina normale può avvenire per mutazioni del promotore o per amplificazione del gene, oppure per mutazioni che causano un incremento dell'attività molecolare specifica della proteina, cioè l'incremento dell'attività molecolare di ogni singola molecola (es. numero di turnover degli enzimi).

l'acquisizione di una nuova proprietà molecolare da parte della proteina codificata
La proteina può acquisire una nuova proprietà molecolare in conseguenza di una mutazione negli esoni del gene, di una fusione del gene con un altro gene o con DNA non codificante, o da incrementi nelle ripetizioni in tandem di triplette di nucleotidi già presenti nel gene o vicino al gene.

Esempi di patologie causate da mutazioni che portano a riduzione/perdita di attività del prodotto genico sono: l'emofilia, la fenilchetonuria, l'albinismo, distonia sensibile alla DOPA, la Met-emoglobinemia (figure E-3 ÷ E-7). Un esempio di patologia dipendente da mutazioni che provocano l'incremento dell'attività del prodotto genico è il contributo dato alla cancerogenesi dagli oncogeni.
Esempi di patologie causate da mutazioni che conferiscono nuove proprietà al prodotto genico sono: l'anemia delle cellule falciformi, causata da una mutazione che conferisce alle subunità-β dell'emoglobina la nuova proprietà di far polimerizzare l'Hb deossigenata. Ciò altera la struttura delle cellule che a sua volta causa la riduzione della loro vita con conseguente anemia; inoltre le cellule tendono ad intasare i capillari. La proteina α-1-antitripsina è un inibitore naturale dell'attività dell'enzima proteolitico elastasi. Una sua forma mutata lega ed inibisce la trombina, causando una malattia emorragica letale. Altri esempi sono il gene chimerico della leucemia mieloide cronica (cromosoma Filadelfia, figura E-2), alcune talassemie-α causate da mRNA o subunità-α instabili. Nella talassemia-α (HB Constant Spring) il codice di stop è mutato e la traduzione continua fino ad un successivo stop. La proteina più lunga è instabile.

Patologie umane causate da triplette di nucleotidi ripetute in tandem

Delle 64 triplette di nucleotidi (sequenze di tre nucleotidi) che si possono ottenere da combinazioni delle 4 basi, molte si ritrovano ripetute nei microsatelliti umani (Tabella 1-2). Alcuni tandem delle triplette: CAG, CGG, CTG si trovano in prossimità o all'interno di geni e, se durante la replicazione del DNA il numero delle loro ripetizioni aumenta (figura 3-17), possono causare patologie. Le ripetizioni delle tre triplette sono presenti anche nei soggetti normali, tuttavia oltre un certo numero di ripetizioni (diverso da gene a gene) si ha l'insorgenza della patologia. Per alcuni geni esiste una condizione intermedia detta "pre-mutazione" (Tabella E-1 e figura E-8). Le ripetizioni al di sotto di un certo numero sono stabili in meiosi e mitosi, mentre oltre un valore soglia sono trasmesse variate (incrementate o ridotte) dai genitori ai figli. In genere c'è la tendenza ad aumentare.

Meccanismi molecolari delle alterazioni geniche provocate dai tandem di triplette di nucleotidi

1. Ripetizioni della tripletta CAG nella parte codificante del gene introducono sequenze di poliglutamina nella proteina codificata. La trascrizione e la traduzione sono normali. La malattia di Huntington (HD) è particolare perché il relativo gene causa HD se mutato da un certo numero di triplette CAG, mentre se distrutto da traslocazioni non causa HD né altre malattie.
2. Ripetizioni al 5' non codificante (5'UTR) causano alterazioni nella metilazione e nella struttura della cromatina di geni adiacenti inibendo la loro trascrizione. L'alterazione della cromatina del cromosoma X è indicata come "fragile X" (FRX).
3. Ripetizioni al 3'UTR (per ora un solo caso) del gene della proteina cinasi (gene DMK = Distrofia Miotonica Cinasi) causa la distrofia miotonica. La trascrizione del gene non è alterata, né la struttura del prodotto genico. Si ipotizza che triplette ripetute alterino la struttura della cromatina e provochino alterazioni in altri geni.

Mutazioni di uno stesso gene possono causare fenotipi patologici diversi

In relazione al tipo di mutazione per alcuni geni si ha perdita o incremento di attività molecolare, o acquisizione di una nuova proprietà molecolare da parte della proteina codificata. All'interno di questi tre tipi di mutazione si possono avere patologie diverse in relazione a mutazioni diverse dello stesso gene. In alcuni casi di mutazione, in relazione alla quantità persa di attività molecolare si manifestano due o più patologie (effetto dose). Ciò può verificarsi perché il gene è espresso in cellule diverse e la proteina espressa, pur avendo la stessa attività molecolare (es. la stessa attività catalitica), nei vari tipi cellulari svolge funzioni fisiologiche diverse richiedenti valori diversi di attività molecolare. La perdita parziale dell'attività molecolare e quindi della funzione fisiologica cellulare può risultare patologica nelle cellule che necessitano un'alta attività

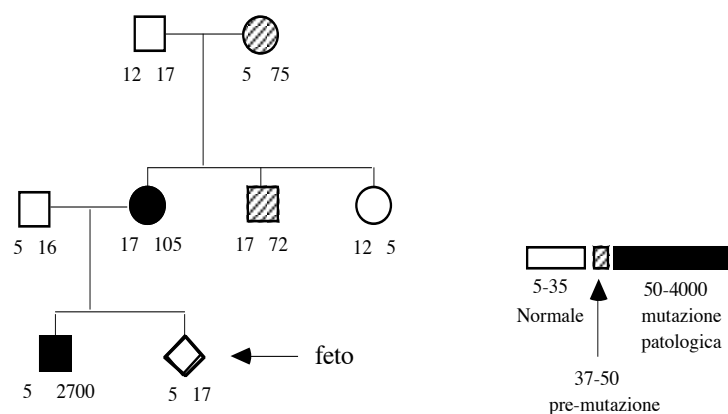



Figura E-8 Contenuto di ripetizioni della tripletta CTG nel locus della distrofia miotonica nei membri normali e malati di una stessa famiglia. I numeri indicano il numero delle ripetizioni della tripletta CTG in ciascun allele. In questa famiglia le ripetizioni delle pre-mutazioni hanno valori superiori alla media.

Tandem di triplette di nucleotidi instabili nel genoma umano*

Patologia	Locus del gene	Localizzazione delle ripetizioni nella regione	sequenze ripetute	Lunghezza normale	Pre-mutazione	Mutazione completa
Malattia di Huntington	4p16.3	Codificante	(CAG)n	9-35	?	37-100
Malattia di Kennedy	Xq2ì	Codificante	(CAG)n	17-24	-	40-55
Atassia spino-cerebellare (SCA1)	6p23	Codificante	(CAG)n	19-36	?	43-81
Atrofia dentato-rubropallidoluysiana (DRPLA)	12p	Codificante	CAG)n	7-23	?	49->75
di Machad-Joseph (MJD, SCA3)	14q32.1	Codificante	(CAG)n	12-36	?	67->79
Cr. X Fragile sito A >1000 (FRAXA)	Xq27.3	5' UTR	(CGG)n	6-54	50-200	200-
Cr. X Fragile sito E (FRAXE)	Xq28		(CCG)n	6-25	?	>200
Cr. X Fragile sito F (FRAXF)	Xq28	?	(GCCn	6-29	?	>500
Cr. 16 Fragile sito A (FRA 16A)	16q22	?	(CCG)n	16-49	-	1000-2000
Distrofia miotonica (DM)	19q13	3' UTR	(CTG)n	5-35	37-50	50-4000

Tabella E-1. (* da Human Molecular Genetics, Strachan and Read, Bios Scientific Publisher, 1996 parzialmente modificato).

molecolare ma non in altre cellule dove è sufficiente una bassa attività molecolare per mantenere la normale fisiologia cellulare. In relazione al grado di perdita di attività molecolare si hanno sintomi clinici diversi (patologie diverse). Alcuni esempi. Le talassemie-α sono causate da mutazioni dei geni delle subunità α del Hb (esistono 2 geni α identici su ciascun cromosoma 16 umano) provocando un effetto dose da riduzione della funzione.
I soggetti normali sono $\alpha\alpha/\alpha\alpha$; soggetti α-/α- oppure $\alpha\alpha$/-- manifestano sintomi leggeri; soggetti α -/-- sono malati gravi; soggetti che non hanno geni α (--/--) non sono vitali e muoiono per "idrope fetale letale" nel corpo materno o poco dopo la nascita. Un altro esempio di fenotipo patologico sensibile alla dose dell'espressione del relativo gene è dato dai livelli decrescenti dell'enzima ipoxantina-adenina fosforibosil-transferasi (HPRT) che causano fenotipi patologici diversi (Tabella E-2).
Altri esempi: il gene RET se mutato dominante negativo causa la malattia di Hirschprung (assenza di gangli nervosi mioenterici e della submucosa del retto) -

% attività catalitica HPRT	Fenotipo
>60	normale
8 - 60	Iperuricemia (gotta), neurologicamente normale
1,6 - 8	Disturbi neurologici (corea-atetosi: movimenti involontari scomposti delle mani, braccia e testa)
1,4 - 1,6	Sindrome Lesch-Nyahan (automutilazioni, corea-atetosi, ma intelligenti)
<1,4	Sindrome Lesch-Nyahan classica (automutilazioni, corea-atetosi e ritardo mentale).

Tabella. E-2. (da Human Molecular Genetics, Strachan and Read, Bios Scientific Publisher, 1996 parzialmente modificata).

mentre con mutazioni che portano alla sostituzione di specifici aminoacidi causa (per acquisizione di nuove funzioni) tumori della tiroide ed altri tumori di cellule endocrine; il gene PMP22 codifica una proteina della mielina periferica, mutazioni (puntiformi o delezioni) che ne riducono l'espressione causando una neuropatia con paralisi pressorie, mentre l'eccesso di espressione dello stesso gene causa una neuropatia motoria e sensoria (malattia di Charcot-Marie-Tooth) (probabilmente le due patologie risultano da una non corretta proporzione dei componenti la mielina periferica); mutazioni diverse del gene del recettore degli androgeni (AR) causano patologie diverse: mutazioni che fanno perdere l'attività molecolare di AR causano insensibilità agli androgeni (sindrome della "femminizzazione" testicolare) mentre l'inserimento di triplette CAG nel gene

strutturale (probabilmente per acquisizione di una nuova attività molecolare) causa la malattia di Kennedy (atrofia spinobulbare muscolare).
L'enzima NAD citocromo-b5 riduttasi è codificato da un gene che si trova sul cromosoma umano 22. L'enzima, oltre che nei globuli rossi dove è presente in forma solubile, si trova associato alla membrana del reticolo endoplasmatico di molti tessuti. Nei globuli rossi l'enzima riduce l'Hb ossidata (figura E-7). L'enzima, nelle cellule nucleate che daranno origine ai globuli rossi, si trova associato al reticolo endoplasmatico e con la maturazione delle cellule diviene solubile perché perde un terzo della sua molecola. Il peptide perso è idrofobico e rappresenta la parte che ancora l'enzima alla membrana del reticolo endoplasmatico nei precursori dei globuli rossi. Nelle cellule dei tessuti, la proteina enzimatica non perde il peptide idrofobico e rimane ancorata alla membrana del reticolo endoplasmatico. Nei tessuti, l'enzima svolge una funzione fisiologica diversa, catalizza le reazioni di deidrogenazione degli acidi grassi (rendendoli insaturi) e partecipa a reazioni di idrossilazione degli steroidi e di composti esogeni, farmaci inclusi.
Da mutazioni diverse dello stesso gene che codifica l'enzima citocromo-b5 riduttasi si hanno due forme di Met-emoglobinemia (recessiva). Esse sono per un aspetto simili ma per un altro molto diverse.
In una forma (non letale) si ha solo Met-emoglobinemia, nell'altra (10% dei pazienti), oltre la Met-emoglobinemia, si hanno disturbi a carico del sistema nervoso centrale e morte prematura.
Si ritiene che la forma letale risulti dalla perdita totale dell'attività catalitica come per delezione del gene, oppure che la mutazione interessi sia la parte catalitica che la parte idrofobica dell'enzima. Per ipotesi, se l'enzima subisce una mutazione che ne riduce l'attività catalitica, ma non il legame al reticolo endoplasmatico, si osserverebbe il danno solo a livello dei globuli rossi, poiché si assume che i globuli rossi, a causa della grande quantità di molecole di emoglobina con le quali devono interagire, hanno bisogno di una attività enzimatica maggiore di quella necessaria alle cellule nervose. Mentre, se la mutazione riduce l'attività catalitica ed inoltre non permette l'ancoraggio al reticolo endoplasmatico, ciò causerebbe la perdita di funzione anche a livello del sistema nervoso perché nelle cellule nervose l'enzima sarebbe obbligato ad avere una specifica localizzazione subcellulare (la membrana del reticolo endoplasmatico), non necessaria per i globuli rossi.

Sensibilità ai fattori ambientali nei portatori di alleli patologici recessivi

Esistono patologie monogeniche recessive che rendono i portatori molto sensibili agli effetti dei fattori ambientali ed in particolare dei farmaci. In questi casi la mutazione, oltre ad essere responsabile in omozigosi della patologia, in eterozigosi è responsabile della sensibilità ad un fattore esogeno (ambientale o alimentare). Questa sensibilità può essere molto insidiosa perché il composto esogeno può essere anche un farmaco.

Esempi:
1. Sulfamidici ed analgesici possono causare anemia emolitica in portatori eterozigoti anche se asintomatici della mutazione responsabile della carenza dell'enzima G6P-deidrogenasi dei globuli rossi (recessiva legata al cromosoma X).
2. Il fumo delle sigarette può avere effetti deleteri sui portatori omozigoti (e forze anche eterozigoti) della deficienza della proteina $\alpha-1$ antitripsina (inibitore di molte proteasi) che predispone all'enfisema polmonare.
La maggiore sensibilità a fattori ambientali in individui portatori di alleli mutati può essere spiegata: il fattore ambientale altera una o più proteine normali che sono i substrati naturali dell'enzima mutato ed in questo modo aggrava l'alterazione causata dalla mutazione. Questo è il caso delle mutazioni che portano a riduzione dell'attività G6P-deidrogenasica e quindi ad una minore sintesi di NADPH che nei globuli rossi è utilizzato per eliminare composti ossidanti che si formano nella cellula e per mantenere lo stato ridotto di alcune proteine cellulari inclusa l'emoglobina (figura E-7). L'introduzione nell'organismo di composti come nitriti e nitrati o di farmaci come i sulfamidici favorisce l'ossidazione dell'emoglobina che provoca un incremento nella richiesta di coenzimi ridotti (NADH e NADPH) che non può essere soddisfatta data la carenza dell'enzima G6P-deidrogenasi mutato e che finisce per provocare anemia.

Fenocopie di malattie genetiche

Fenocopia di una malattia genetica è l'alterazione del fenotipo provocata da fattori ambientali o nutrizionali che porti alla formazione di un soggetto con caratteristiche identiche o molto simili a quelle di un soggetto portatore di una alterazione genetica.
Un esempio classico di fenocopia di malattia genetica è dato dalla carenza alimentare di vitamina D durante la crescita di un individuo che provoca manifestazioni molto simili (alterazione nelle ossa, bassa statura ed altre alterazioni anche biochimiche) al rachitismo genetico detto "resistente alla somministrazione di vitamina D" (cioè che non è completamente rimosso/ non completamente curabile con somministrazione di vitamina perché il soggetto è incapace ad assumerla).
Nitrati, derivati dell'anilina e sulfamidici accelerano l'auto-ossidazione del Hb normale fino a causare cianosi come nei portatori di Met-emoglobinopatie genetiche. Questo si verifica perché la normale attività dell'enzima Met-Hb-riduttasi risulta insufficiente a rigenerare l'eccesso di Hb ossidata provocato dall'azione dei composti sopra indicati. Questo effetto è drammaticamente più forte nei portatori (etero ed omozigoti) di Met-Hb-riduttasi mutata. I portatori eterozigoti di una patologia recessiva (senza sintomatologia), essendo mancanti della funzione di un allele, rispetto agli individui con ambedue gli alleli normali, mostrano una sensibilità maggiore ai fattori ambientali ed alimentari e quindi a

manifestare i sintomi della patologia. Egualmente i portatori eterozigoti di G6P-deidrogenasi (senza sintomi) se esposti ad ossidanti manifestano una anemia emolitica come i portatori omozigoti dello stesso gene mutato. La maggiore sensibilità allo stress ambientale dei portatori di patologie recessive e le conseguenti manifestazioni patologiche sono casi particolari di fenocopie perché misti (genetico-ambientali). Le fenocopie indicano che le malattie causate da un gene mutato che esprime una proteina inattiva (o che non la esprime affatto), possono essere mimate da carenze alimentari di un composto essenziale come le vitamine o da molecole esogene che introdotte nell'organismo (per inalazione, contatto, alimentazione, ecc.) legandosi alla proteina prodotta dal gene normale ne inibiscono l'attività molecolare. Ciò indica che le alterazioni patologiche sono conseguenti al non funzionamento di una proteina perché alterata nella concentrazione e/o nell'attività per carenza del suo substrato o cofattore.

Ogni problema che ho risolto è diventato una regola che successivamente è servita a risolvere altri problemi.
René Descartes

Appendice F

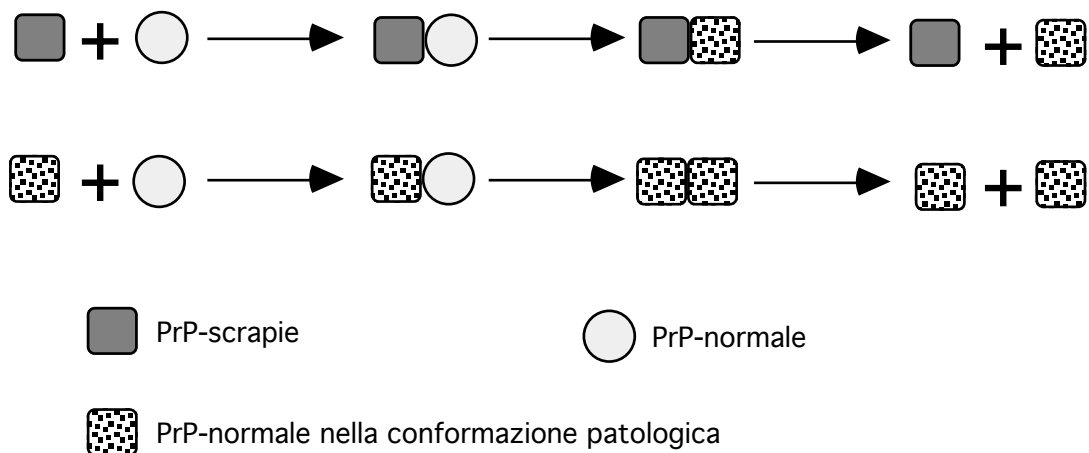
Prioni, una nuova classe di agenti infettivi ed una nuova patologia genetica[27]

I prioni (prions, proteinaceous infectious particles) patologici sono proteine mutate prodotte nell'encefalo od introdotte con gli alimenti che causano encefalopatie spongiformi mortali inducendo cambiamenti di conformazione in proteine cellulari identiche o molto simili a loro (prioni normali). Scrapie è una encefalopatia spongiforme, diagnosticata gia nel 1700, che colpisce pecore e capre. Gli animali che ne sono colpiti oltre a manifestare perdita di coordinazione ed irritabilità, hanno un forte prurito che li spinge a strusciarsi ripetutamente a staccionate od altro fino grattar via (scrape) la lana dalla loro pelle. L'encefalopatia spongiforme colpisce anche altri animali (visone, alce ed uomo), ha tempi lunghi di incubazione (nell'uomo anche decine di anni). Essa è detta spongiforme perché provoca la formazione di cavità nel cervello rendendolo simile ad una spugna.

Una encefalopatia spongiforme riscontrata nei membri di alcune tribù della Papua-Nuova Guinea è detta kuru (causa atassia seguita da demenza). Essa era trasmessa all'interno delle tribù in conseguenza del cannibalismo attuato come rito. Per onorare il parente morto ne veniva mangiato il cervello. Una volta che i membri delle tribù sono stati convinti a cessare il cannibalismo è cessata anche la malattia. La forma di encefalopatia spongiforme che è riscontrata in tutto il mondo è detta malattia di Creutzfeldt-Jacob, è mortale, si manifesta come demenza, colpisce una persona per milione (in genere all'età di 60 anni), nella maggior parte dei casi è sporadica (non trasmessa geneticamente), mentre nel 10-15% dei casi è ereditaria. Rari casi sono iatrogeni (causati da terapie per altre patologie) come può avvenire nei trapianti di cornea. Altre due patologie ereditarie mortali sono dipendenti da prioni: la malattia di Gerstmann-Straussler-Scheinker (che si manifesta come atassia per danni al cervelletto) e l'insonnia familiare mortale (demenza conseguente all'insonnia, scoperta da ricercatori italiani: Elio Lugaresi, Rossella Medori e Pierluigi Gambetti). Si ipotizza che altre patologie neurodegenerative come la malattia di Alzheimer (ed anche di patologie che interessano la muscolatura) possono essere causate da prioni diversi da quelli responsabili delle encefalopatie spongiformi.

I prioni (PrP) non patogeni sono normali costituenti delle cellule nervose dei mammiferi, sono anche detti prioni cellulari, e la loro attività molecolare è ignota. I prioni che causano encefalopatie sono prioni che portano mutazioni puntiformi e sono detti prioni dello scrapie (PrP-scrapie); con questo termine sono indicati anche i prioni che causano encefalopatie simili allo scrapie in animali e nell'uomo. Nell'uomo, il prione patologico risulta da 18 mutazioni puntiformi diverse che inducono la proteina ad assumere una conformazione

terziaria particolare. Il cambiamento di struttura da PrP normale a PrP-patologico consiste nel cambiamento di 4 tratti di struttura in alfa-elica (prione normale) in tratti di struttura-beta (prione patologico). A conferma di ciò, le mutazioni umane del prione si trovano all'interno o sui bordi delle quattro strutture in alfa-elica convertibili in strutture-beta. La struttura-beta (patologica) rende il PrP insolubile, resistente alle proteasi ed alle temperature di 360°C.
La struttura-beta è assunta spontaneamente nei PrP mutati (prioni-scrapie), mentre è indotta nei prioni normali dall'associazione con PrP-scrapie che in qualche modo (alimentazione e/o via ematica, diffusione) entrino in contatto con loro. La conversione da PrP normale in patologico è stata dimostrata anche *in vitro* mescolando le due forme di PrP. La progressione della patologia (patogenesi) è di tipo "a cascata" perché anche i PrP normali che hanno assunto la conformazione patologica hanno la capacità di indurre la conformazione patologica in altri PrP normali rendendo così il meccanismo molecolare irreversibile. Nella patologia trasmessa geneticamente, le proteine mutate neosintetizzate probabilmente non adottano subito la conformazione patologica, altrimenti la malattia si manifesterebbe nei bambini, mentre deve passare molto tempo prima che la prima molecola assuma la conformazione patologica e poi ne deve passare ancora molto prima che le molecole patologiche si accumulino e danneggino le cellule nervose fino a causare i primi sintomi. I PrP-scrapie si accumulano nei lisosomi e si ipotizza che ne provochino la rottura. Gli enzimi litici (proteasi e nucleasi) fuoriusciti causerebbero la distruzione delle cellule nervose creando i buchi osservati nell'encefalo dei pazienti. I frammenti di PrP liberati dai lisosomi sarebbero responsabili delle placche osservate nell'encefalo di alcuni pazienti. Il meccanismo della patogenesi della encefalopatia spongiforme sporadica (da mutazioni somatiche) si assume identico a quello della patologia trasmessa geneticamente o per introduzione nell'organismo di PrP-scrapie.
Nell'encefalopatia spongiforme sporadica, il PrP-scrapie viene sintetizzato in una unica cellula nervosa e con la degenerazione della cellula, i PrP-scrapie raggiungono (direttamente o per via ematica) altre cellule nervose, ed associandosi ai PrP normali in esse contenuti, le inducono ad assumere la conformazione patologica. In questo modo, la patologia si diffonde a nuove cellule e queste poi la diffondono ad altre ancora (questo meccanismo ha aspetti di una infezione e di una metastasi molecolare). Come sopra indicato l'encefalopatia spongiforme può essere trasmessa per infezione di prioni-scrapie introdotti nell'organismo con l'alimentazione o per via ematica (lesioni nelle mani o nelle pareti delle mucose della bocca, stomaco o intestino). Si assume che il PrP-scrapie raggiunga le cellule nervose inducendo i prioni normali ad assumere la conformazione patologica e quindi a diffondere la patologia. Si può anche ipotizzare che i prioni normali, se associano particolari agenti inquinanti o farmaci (introdotti per curare altre patologie), possano da questi essere indotti ad assumere la conformazione patologica determinando poi la reazione a catena che porta ad instaurare l'encefalopatia. Tuttavia esiste anche la speranza di

poter curare le encefalopatie sintetizzando farmaci capaci di indurre i PrP-scrapie ed i PrP normali ad assumere/mantenere la conformazione normale o mediante vaccinazioni anti-PrP-scrapie. La barriera di specie, cioè la difficoltà che il PrP-scrapie di una specie trova nell'instaurare la patologia in un'altra specie, è in relazione diretta al numero di aminoacidi diversi presenti nelle sequenze aminoacidiche dei PrP appartenenti a specie diverse. Ne consegue che maggiore è l'identità di sequenza (% di aminoacidi identici) esistente tra PrP di specie diverse, maggiore è la probabilità della trasmissione della patologia tra le due specie. I prioni di ovini e bovini differiscono di solo 7 aminoacidi e ciò può spiegare come prioni-scrapie di ovini contenuti in farine alimentari abbiano infettato molti bovini. Il passaggio per via alimentare o ematica di PrP-scrapie bovini all'uomo dovrebbe essere più difficile perché i relativi prioni differiscono per più di 30 aminoacidi. Tuttavia, di recente in Gran Bretagna si è osservato l'insorgere nell'uomo di circa 30 casi di encefalopatie, varianti della malattia di Creutzfeldt-Jacob. Queste varianti della patologia sono state poste in relazione alla più estesa epidemia dei bovini (circa 180.000 casi) le cui carni inizialmente sono state messe in commercio. L'inefficienza della barriera di specie alle infezioni di PrP-scrapie è stata poi dimostrata instaurando la patologia in topi transgenici in cui era stato transfettato il gene del PrP-scrapie omologo o di specie diverse oppure iniettando PrP-scrapie di una specie nell'encefalo dell'altra.

Il PrP-scrapie associandosi al PrP-normale lo induce ad assumere la conformazione patologica. Il PrP-normale non solo mantiene questa conformazione ma è anche capace di indurla in altri PrP-normali.
La particolarità del meccanismo di propagazione (a cascata) rende la patologia progressiva ed irreversibile.

Topolini (Mus musculus) in cui è stata provocata mediante knockout la distruzione omozigote del gene PrP risultano normali per almeno 70 settimane (circa metà della loro vita) e risultano tali anche dopo aver iniettato molecole di PrP-scrapie nel loro encefalo. Gli esperimenti hanno fornito alcuni dati importanti:
- il PrP non è importante per la prima parte della vita dei topi;
- il PrP-scrapie iniettato non riusciva a causare la patologia se non era presente quello normale prodotto dall'animale, probabilmente perché, mancando il PrP endogeno, il PrP-scrapie iniettato (data la sua quantità limitata) non poteva instaurare la patologia;
-che l'instaurarsi della patologia era causato dal PrP-scrapie che modificava la conformazione del PrP presente nelle cellule (e non da un virus molto cercato e mai trovato). Tuttavia, poiché in tempi lunghi (circa 90 settimane) la carenza (per knockout) di PrP causa problemi di tipo neurologico, si è dedotto che il PrP normale deve avere una azione protettiva contro l'encefalopatia (un effetto di questa azione è il mantenimento in vita delle cellule del Purkinje). Poiché l'azione di protezione è persa anche con la trasformazione del PrP in PrP-scrapie, i sintomi causati dalla carenza di PrP devono far parte del quadro di sintomi della encefalopatia spongiforme. Anche l'eccesso di PrP (per transfezione) causa problemi: degenerazione di nervi periferici e dei muscoli.

Stanley B. Prusiner, premio Nobel 1997 per la fisiologia e medicina per la scoperta dei prioni, con i suoi studi ha mostrato che i prioni sono una nuova classe di agenti infettivi, privi di acidi nucleici, diversi da virus, batteri e parassiti sia nel meccanismo di riproduzione e che nel meccanismo di patogenesi. Lo stesso autore suggerisce che i prioni diversi da quelli responsabili delle encefalopatie possono essere responsabili di altre patologie e di casi clinici di cui si conoscono le manifestazioni ma non l'eziologia.
Il meccanismo mediante il quale i prioni-scrapie causano l'encefalopatia introduce nuovi concetti in biologia molecolare:
-l'alterazione della funzione di una cellula operata da una particolare conformazione di una proteina e dalla trasmissione di questa conformazione ad altre proteine mediante formazione di un complesso;
-la trasmissione della nuova confermazione può avvenire tra proteine identiche e simili appartenenti anche a specie diverse.